Journal of Economic Behavior and Organization 210 (2023) 265–287

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Polycentricity and adaptation: A multilevel selectionist approach☆

Alexander Schaefer

New York University, School of Law, 110 W. 3rd St. Room 224, New York, NY 10012 USA

ARTICLE INFO

Article history:
Received 20 August 2022
Revised 4 April 2023
Accepted 6 April 2023

Keywords:
Polycentricity
Group selection
Multilevel selection
Elinor Ostrom
Price equation

ABSTRACT

Polycentric organizations allow highly functional, or "group-beneficial," outcomes to emerge from the myopic behavior of rule-guided individuals. How does polycentricity achieve this feat? Drawing on multilevel selection theory, I argue that polycentric orders support successful outcomes by defining group boundaries and reducing within-group fitness variance relative to between-group variance. The Price equation suggests that, by doing so, polycentric orders facilitate a process of collectively beneficial adaptation, including the capacity to evolve mechanisms for monitoring and punishing rule violators.

1. The puzzle of polycentric functionality

Put succinctly, a polycentric governance structure is one that involves multiple decision-making units, each with authority over a specified, but evolving jurisdiction, which interact in various ways, according to a set of overarching rules (Aligica and Tarko, 2012). This structure typically engenders diverse forms of competition (Stephan et al., 2019); for instance, jurisdictions, which may be territorial or non-territorial, often compete for members.

At the same time, many social scientists have documented the surprising functionality of polycentric governance structures. Their desirable attributes include greater resilience (Carlisle and Gruby, 2019), as well as a greater capacity for coordination (Tarko, 2022), information processing, and adaptation (Andersson and Ostrom, 2008).

Together, these features – decentralized competition and high functionality – produce a puzzle: how does decentralized competition generate beneficial social outcomes? Although invisible hand processes often produce a surprising degree of order, they typically presuppose a certain institutional background. Decentralization and competition do not, in general, guarantee functional social order (Wilson, 2016). Without the right institutions the invisible hand is liable to become an "invisible fist" (Anomaly and Brennan, 2014).[1] Only under the right institutional structure do "invisible hand processes" incentivize local actions that produce global benefits.[2] Why is it, then, that polycentric systems often support high-quality

☆ I would like to thank David Schmidtz, Thomas Christiano, Vlad Tarko, and Justin Bruner for many helpful discussions on the topic of this paper. Peter Boettke, Cameron Harwick, and Abigail Devereaux offered insightful comments when I presented earlier versions of this paper at the 2022 Markets & Society Conference and the 2021 Meeting of the Southern Economic Association. I'd also like to thank Mario Iván Juárez García and Matthew Jeffers for their reliable willingness to discuss premature versions of my arguments.

E-mail address: Schaefer.alexander@nyu.edu

[1] See also Martin and Storr (2008).

[2] Elinor Ostrom expressed this issue by referencing the possibility of "local tyrranies" that might arise in self-governing institutional structures (Ostrom, 2009, 282-3).

https://doi.org/10.1016/j.jebo.2023.04.004
0167-2681/

governance? Put more strongly, how can decentralized competition give rise to beneficial order, rather than repugnant equilibria or disorderly chaos?

This paper advances our understanding of polycentric governance by drawing on multilevel selection theory. This theory is arguably our best tool for understanding the relationship between lower level rationality and emergent functionality in unplanned, decentralized systems. It has already been fruitfully applied to identify precise conditions under which group-beneficial adaptations are likely to emerge (Henrich, 2004), but experts on polycentricity have yet to examine the conceptual connection between multilevel selection theory and theories of polycentricity.[3] In an effort to construct this bridge, the paper begins by laying out the formal framework of multilevel selection theory, encapsulated in two formulas derived from Price's equation (Section 2).[4] It then identifies the crucial features – here called "functionality desiderata" – that a system must exhibit in order to achieve group-level success. Through the lens of this formal framework, the next section (Section 3) shows how various features of polycentric governance support collective functionality. To clarify this abstract treatment, Section 3 also examines a concrete example of polycentric governance: the scientific research community. The multilevel selection framework developed in Sections 2 and 3 illuminates the capacity of polycentric governance structures to evolve rules and norms that benefit communities, while suppressing opportunistic behavior.

2. The price equation and multilevel selection

2.1. The basic price equation

When systems evolve as a result of the myopic choices of their members, this generally results "in a breakdown of group-level functional organization" (Wilson, 38). As economists and biologists have long known, lower-level rationality does not automatically scale; social dilemmas abound, leading to suboptimal outcomes for economies and species. One solution is for social planners to direct the behavior of individuals towards well-defined goals, but this solution works well only under special conditions. In more realistic settings, institutional frameworks must allow for decentralized decision-making, and polycentric governance structures offer an extreme version of such frameworks. Given the high degree of local autonomy, we must wonder how polycentric governance structures are able to secure desirable outcomes at the global level. Individuals within various jurisdictions do not, in general, possess the knowledge or the desire to produce group-beneficial outcomes. Yet, somehow, polycentric political organization often promotes effective governance (Thiel et al., 2019). What features of polycentric organization underlie this surprising capacity?

To understand how a decentralized network of decision-making entities can produce and maintain beneficial social order, it will help to draw on multilevel selection theory. In particular, two versions of the Price equation provide deep insight into the conditions in which such social order will arise and persist.

First, we need to establish some notation:

- z: A (phenotypic) trait of some kind, which can be measured (discretely or continuously) with real numbers.
- z_i: The level of z exhibited by the entity $i \in P = \{1, \ldots, n\}$.
- $\bar{z}$: The average level of z in the P-population.
  - Mathematically, $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$
- Δz_i: The change in the level of z from one entity to its offspring.
  - Taking an average across all entities in P, the expectation of Δz_i is $\mathbf{E}[\Delta z] = \frac{1}{n}\sum_{i=1}^{n} \Delta z_i$
- w_i: The "fitness" or average number of offspring produced by entity i.
  - Similarly to $\bar{z}$, we define the average fitness as $\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$
  - Often we will be interested, not in the absolute fitness of an entity, but in its fitness *relative* to the other entities in the group. For this comparative purpose, we define *relative fitness* of entity i as $\gamma_i = \frac{w_i}{\bar{w}}$.

With this notation in hand, a form of the Price equation can be written as follows:[5]

$$\Delta\bar{z} = Cov(\gamma, z) + \mathbf{E}[\gamma\,\Delta z]. \tag{1}$$

Although simple, (Price Eq.) reveals the key components of evolutionary change. It decomposes the total change in $\bar{z}$, that is, the total change in the average level of the z-trait, into two components: $Cov(\gamma, z)$ and $\mathbf{E}[\gamma\,\Delta z]$. $Cov(\gamma, z)$ denotes the statistical association between γ, relative fitness, and z, the level of the z-trait. When these two variables move together, that is when greater relative fitness is associated with higher levels of z, the covariance will be positive: $Cov(\gamma, z) > 0$. When they are totally unrelated, $Cov(\gamma, z) = 0$. And when they move in opposite directions, e.g. higher levels of z-trait go along with lower relative fitness, the covariance will be negative: $Cov(\gamma, z) < 0$. Therefore, this first term is often identified with the evolutionary force of *selection* (Gardner, 2008).

[3] Wilson et al. (2013) have applied multilevel selection to explain how Elinor Ostrom's eight "design principles" (Ostrom, 2016) support successful outcomes, but do not directly address the relationship between polycentricity and multilevel selection.

[4] The Price equation approach is not the only approach to understanding multilevel selection. Okasha (2006), for example, presents a "contextual approach" sometimes provides a better causal decomposition of the forces at play in multilevel selection. However, the Price approach has proven quite successful at illuminating social evolutionary processes (Turchin, 2011).

[5] The subscript i has been dropped on all terms that occur within an expected value of a covariance operation, since this operation is performed across the entire population. See Appendix A.1 for a derivation with more precise notation.

The second quantity on the right-hand side of Eq. (Price Eq.), $\mathbf{E}[\gamma \Delta z]$, denotes the weighted average of the change in z-levels between parents and offspring. Each entity $i \in P$ exhibits level z_i of the z-trait and produces some number of offspring w_i with a level z_i' of the z-trait. If we want to measure the average transmission rate of the z-trait, it makes sense to take the average change between parents and offspring, but for a more accurate measure, we should also heavily weight those entities with many offspring and discount those with few offspring. $\mathbf{E}[\gamma \Delta z]$ achieves this weighting by multiplying each Δz_i by the relative fitness of entity i, which is $w_i/\bar{w} = \gamma_i$.[6] In short, $\mathbf{E}[\gamma \Delta z]$ gives us a measure of the population's overall *transmission bias* or *copying fidelity*. If all entities in P produce offspring with higher levels of z-trait, then $\mathbf{E}[\gamma \Delta z] > 0$. If they all tend to produce entities with lower levels of z-trait, then $\mathbf{E}[\gamma \Delta z] < 0$. When the transmission is perfect, i.e. $z_i = z_i'$ for all $i \in P$, then $\mathbf{E}[\gamma \Delta z] = 0$.[7] In that case, the only evolutionary force in operation is *selection*.

To achieve some intuition for these terms, let us consider a specific interpretation.[8] In this interpretation, we consider two periods, one where entities compete for resources, and another in which they reproduce. We also assume that the z-trait exerts direct causal force in determining the survival of these entities, thus indirectly affecting the expected number of offspring that entities produce. In this simple scenario, $Cov(\gamma, z)$ tells us how well the z-trait promotes survival in the first period. Some entities will die, some will survive, and $Cov(\gamma, z)$ tells us how much the z-trait has contributed to survival ability. In the second period, when the entities reproduce, $\mathbf{E}[\gamma \Delta z]$ adds a further change by representing the amount of the z-trait inherited by the offspring.

This can be made more concrete with some illustrative examples. The first is a standard biological example, while the second example concerns the cultural trait of adhering to a social norm.

1. Consider a population $N = \{1, \ldots, n\}$ of polar bears. For each polar bear $i \in N$, we can measure the heaviness (and corresponding warmth) of its fur as some number $z_i \in (0, 1)$, where 0 corresponds to no fur at all and 1 represents the heaviest possible coat that a polar bear could physically grow. For a population that inhabits its environment of evolutionary adaptation, we can assume that each $i \in N$ will have z_i close to the optimal level of z, call it z^*. This value is optimal in the sense that it maximizes the survival and reproduction of a polar bear. Now suppose there has been some climatic shift, a new ice age has set in, and every bear in N is now insufficiently insulated against the cold. In this case, polar bears with higher levels of z will survive longer and produce more offspring than polar bears with lower levels. Hence, there will be a positive correlation between the heaviness of a bear's coat, z, and its relative fitness, γ. This means that $Cov(\gamma, z) > 0$ in the relevant range. If the heaviness of a bear's coat is perfectly inherited by its offspring, then this is the end of the story. However, for some genetic reason, the heaviness of fur may not be perfectly inherited. If copying errors tend to produce lighter coats, then $\mathbf{E}[\gamma \Delta z] < 0$. If they tend to produce heavier coats, then $\mathbf{E}[\gamma \Delta z] > 0$. And if the error is symmetrically distributed, then $\mathbf{E}[\gamma \Delta z] = 0$. The Price equation therefore decomposes the evolutionary process into two distinct effects: (1) selection and (2) inheritance.
2. In an article on the evolution of social norms, Ostrom (2014) distinguishes between rational egoists and norm-following cooperators. Rational egoists will maximize their material holdings in any strategic interaction, while norm-following cooperators follow a rule of initiating cooperation when they estimate that others will reciprocate.[9] Let our population $P = \{1, \ldots, m\}$ contain a mix of both types, and let z represent the types so that $z_i = 0$ if agent i is a rational egoist, and $z_i = 1$ if agent i is a norm-following cooperator.[10] To determine which type of agent will have higher relative fitness, γ, we must know something about the type-distribution within the community of interaction. As Ostrom (2014, 243-4) tells us, following norms will be advantageous "so long as almost everyone reciprocates. If a small group of users identify each other, they can begin a process of cooperation." This assumes that cooperators will be able to form a network of cooperators, excluding rational egoists. If we assume that materially successful agents have an advantage in passing on their norms, then in a population with this sort of network structure, $Cov(\gamma, z) > 0$. However, in a different network structure, one where there is no way to exclude rational egoists, norm-following cooperators will likely fall prey to opportunistic exploitation, and $Cov(\gamma, z) < 0$.[11] Like in the polar bear example, if norms can be taught or inherited with perfect accuracy, or if the errors are the same for both traits, then this is the end of the story. If we suppose, on the other hand, that the cooperative norm is harder to teach accurately, then $\mathbf{E}[\gamma \Delta z] < 0$. But if it's easier to teach accurately, then $\mathbf{E}[\gamma \Delta z] > 0$. As in the biological example, the Price equation again decomposes the evolutionary process into the distinct effects of selection and copying error.

This decomposition seems rather straightforward: some part of evolution will be due to the relation between fitness and the z-trait, and some part of it will be due to the ability of entities to actually pass the z-trait on to their offspring. We might think of the first term, $Cov(\gamma, z)$, as representing the basic evolutionary idea that if a trait helps an entity survive

[6] A simpler, but slightly more technical, way of understanding $\mathbf{E}[\gamma \Delta z]$ construes it as the expected value of random variable Δz whose probability distribution assigns a probability of $w_i/\bar{w}$ to each Δz_i.

[7] This also occurs when copying errors are random and symmetric about the mean.

[8] This is Okasha's "temporal interpretation" (Okasha, 2006, 24).

[9] Ostrom (2014, 238) also introduces a third type: "willing punishers." I set these aside until the discussion of punishment in Section 3.2.

[10] Notice that the mathematics of the Price equation operate similarly for discrete traits, like this one, and for continuous traits, as in the polar bear example.

[11] Again, we are ignoring the meta-norm of punishment until later in the paper (Section 3.2)

and reproduce, it will proliferate. The second term, $\mathbf{E}[\gamma \Delta z]$, adds the obvious qualification that there are sometimes copying errors between generations. The offspring do not inherit the exact level of z exhibited by their parents.

The standard way of writing the Price equation takes Eq. (Price Eq.) and multiplies through by $\bar{w}$ to yield:

$$\bar{w}\Delta\bar{z} = Cov(w, z) + \mathbf{E}[w\Delta z] \tag{2}$$

Notice that the only real change is that γ has been replaced with w. This is just because $\gamma_i = w_i/\bar{w}$, and we multiplied through by $\bar{w}$. From this basic form of the Price equation, we can derive a simple expression that models multilevel selection.

2.2. The cultural, multilevel price equation

The Price equation purports to be an entirely general description of any evolutionary process (Frank, 1998, 13). In theory, then, it should provide a way of modeling the process of cultural, multilevel selection, e.g. institutional evolution within a polycentric political framework. To see this, we first need to offer an interpretation of Eq. (2). Suppose we have a set of groups indexed by $j \in \{1, \ldots, N\}$. For any individual i in group j, let $z_i \in \{0, 1\}$ represent adherence to a rule or institutional feature that is "altruistic" in a technical sense. That is, z_i represents adherence (1) or non-adherence (0) to a rule that promotes the fitness of the group j (defined as the average individual fitness within group j), but which decreases that of the individual i within group j. Accordingly, $z_j \in (0, 1)$ will represent the average level of the altruistic trait exhibited by individuals within group j, and the variable w_j indicates the average *fitness* within group j. In other words w_j indicates the number of people that are "influenced" by an average member of group j.

In the cultural version of the Price equation, the notion of "influence" replaces that of reproduction. This is because (unlike DNA) rules, norms, ideas, and other cultural replicators can spread without organisms producing offspring. Though having more offspring may lead to more copies of a rule, increased copying of a rule does not entail increased offspring. Instead, w_j measures the average number of other individuals that will copy a member i of group j by imitating group i's rule level z_i. This marks a major departure from models of genetic evolution, and for the remainder of this paper, it will be crucial to keep in mind that "fitness" refers to *cultural* fitness – i.e. the ability to influence others in their selection of rules of conduct.

Another difference between biological evolution and cultural evolution concerns the *criterion of selection*. Setting aside intentional domestication and attempts at eugenic planning, selection in the biological realm occurs without foresight or conscious intent. Mutations happen randomly, and an organism survives and produces offspring without regard to the effect it has on population genetics. A genetic sequence is the byproduct of choices made without regard to their effects on the genome. By contrast, as Elinor Ostrom emphasizes when discussing institutional evolution, "[i]nstead of blind variation, human agents...use reason and persuasion in their efforts to devise better rules" (Ostrom, 1999, 524). Thus, cultural variation and selection involves conscious, rational choice. While random chance does play some role – "the process of choice always involves experimentation" (Ostrom, 1999, 524) – we often change our rules because a different rule seems to work better than the one that's currently in place.

With these two differences in mind, consider an example to illustrate the process of cultural evolution.[12] Let $z_i = 1$ represent i adhering to a norm of working hard in group endeavors. As an individual, i could improve her situation by violating this rule and choosing $z_i = 0$, a life inspired by the "beach boys" or the "flower children" Buchanan (1994, 28). Adherence to the work ethic benefits the members of i's group, and an unwillingness to adhere degrades their quality of life. Consider two groups, j and k. Group j mostly adheres to the work ethic norm, only 20% of its members are beach bums ($z_j = .8$). Group k rejects the importance of work ethic; only 20% of its members feel an obligation to work hard ($z_k = .2$). Recalling the Elinor Ostrom's point that cultural evolution is often driven by conscious choice, suppose that each individual follows a policy of "success bias" or copying norms that seem to promote success.[13] The payoff to an industrious individual ("altruistic") i in group $m = j, k$ will be denoted u_a^m, while the payoff of shirking ("egoist") will be denoted u_e^m. Because laziness confers a personal benefit, while hard work involves foregoing this benefit, these utility functions are defined as follows:[14]

$$u_a^m = az_m - cz_i = az_m - c$$
$$u_e^m = az_m$$

Here, a is a scalar that measures the benefit of increased altruism z within one's group, m. By contrast, c is a scalar that represents the personal cost to agent i of increasing her level of altruism (work ethic). Because we are treating altruism as a binary trait, we have $z_i = 1$ for the hard worker i and $z_j = 0$ for the shirker j.

Will the work ethic spread or recede? At time t, let $\dot{p}_a^t$ represent the rate of change in the proportion of individuals who adhere to the altruistic work ethic norm, and $\dot{p}_e^t$ represent the rate of change in the proportion of those who are egoistic,

[12] This example is inspired by Buchanan (1994, 5–31), who argues that the cultural norm of "work ethic" is altruistic in a way that maps onto the technical sense in which altruism is defined here.

[13] This is a form of social learning that allows modeling in terms of a "replicator dynamic." For an explanation and derivation of the replicator system below, see Gintis (2009, 270–3).

[14] It's no coincidence that the form of the utility function in this simple example resembles that employed in standard models of public goods. See, for example, Samuelson (1954).

lazy, free-riders. Modeling the problem in terms of simple replicator dynamics yields the following equations:

$$\dot{p}_a^t = p_a^t(u_a^t - \bar{u}^t)$$
$$\dot{p}_e^t = p_e^t(u_e^t - \bar{u}^t)$$

where u_a is the average payoff to hard workers, u_e is the average payoff to shirkers, and p_m^t denotes the fraction of the total population adhering to strategy $m = a, e$ at time t.

When will work ethic spread, i.e., when will we have $\dot{p}_a^t > 0$? Only when $u_a^t > \bar{u}^t > u_e^t$. At the initial state, $t = 0$, we have

$$u_a = .8[a(.8) - c] + .2[a(.2) - c] = .68a - c$$
$$u_e = .2[a(.8)] + .8[a(.2)] = .32a$$

Thus, in this initial state, the work ethic norm will be drawing more adherents if and only if...

$$.68a - c > .32a$$
$$.36 > \frac{c}{a}$$

In other words, if the benefits of *living in a group of hard workers* outweighs the personal cost of adhering to a demanding work ethic, then work ethic will tend to spread.[15]

The basic reasoning governing this simple example can be represented analytically by another version of the Price equation:[16]

$$\bar{w}\Delta\bar{z} = Cov(z_j, w_j) + \mathbf{E}\left[w_j \Delta z_j\right] \tag{3}$$

$$\bar{w}\Delta\bar{z} = Cov(z_j, w_j) + \mathbf{E}\left[Cov(w_{ij}, z_{ij}) + \mathbf{E}\left[w_{ij}\Delta z_{ij}\right]\right] \tag{MLPE}$$

The addition of the subscript j in Eq. (3) indicates that we are considering *group* quantities: z_j is the average level of trait z within the group and w_j is the average fitness of the group. The addition of the subscript i in the next line, (MLPE) indicates that we are considering within-group covariances and expectations, taken over individuals within a fixed group j. As above, the term $\mathbf{E}\left[w_{ij}\Delta z_{ij}\right]$ expresses non-selective forces, such as transmission bias or copying error. To focus on evolutionary forces we will ignore this term by setting it equal to 0.

The Price equation represents the proliferation of a particular trait in a population, but it's important to bear in mind that the fitness of a particular trait will depend upon which other traits are prevalent or absent in the population. For instance, the deadliness of a cobra's venom enhances fitness only to the extent that the cobra also has sharp teeth capable of piercing animal flesh. The same idea applies, perhaps even more strongly, to cultural traits. The most important example in this regard is that of punishment. Consider a norm of respecting property rights, z. In the absence of punishment, an opportunistic free rider in group j may be better off disregarding this norm: $\mathbf{E}\left[Cov(w_{ij}, z_{ij})\right] < 0$. If, on the other hand, there is a commonly accepted norm of punishing those who violate property rights, then accepting the norm of property may prove advantageous: $\mathbf{E}\left[Cov(w_{ij}, z_{ij})\right] > 0$. In general, punishment can transform an otherwise altruistic trait into one that agents must adhere to for the sake of their own success. Insofar as punishment is costly, however, the norm of punishment itself is altruistic. The crucial issue of punishment will be discussed further in Section 3.2.

The next subsection presents some important implications of the cultural, multilevel version of the Price equation. In particular, two equations derived from (MLPE) provide a clean representation of the conditions under which group beneficial rules are likely to evolve.

2.3. Implications of the price equation

One more mathematical fact is required to complete the derivation. If we denote as β_1 the regression coefficient for w_j on z_j and β_2 the regression coefficient for w_{ij} on z_{ij}[17], then we have:

$$Cov(z_j, w_j) = \beta_1 Var(z_j)$$
$$Cov(z_{ij}, w_{ij}) = \beta_2 Var(z_{ij}).$$

Letting $\mathbf{E}\left[w_{ij}\Delta z_{ij}\right] = 0$ for reasons mentioned above and substituting these two equations into (MLPE) produces the following:

$$\bar{w}\Delta\bar{z} = \beta_1 Var(z_j) + \beta_2 Var(z_{ij}) \tag{4}$$

[15] More generally, if p_j^t is the fraction of altruists in group j at time t and p_k^t is the fraction of altruists in group k at time t, then altruism will have a higher payoff whenever $\frac{c}{a} < \frac{2(p_j^t)^2 + 2(p_k^t)^2 - p_j^t - p_k^t}{p_j^t + p_k^t}$.

[16] See Appendix A.2 for a derivation of (MLPE).

[17] See Appendix A.3 for a more thorough explanation.

Because they are regression coefficients, β_1 gives a measure of how changing the average level of z within a group affects the average fitness of the group, while β_2 measures how changing the level of z of an individual within a group will change that individual's fitness. The expression (4) thus represents the evolution of the average level of trait z as a composition of a between group portion and a within-group portion. From (4), we can infer that the trait z will spread when...

$$\frac{Var(z_j)}{Var(z_{ij})} > \frac{-\beta_2}{\beta_1} \qquad (*)$$

Assuming that the trait z is "altruistic" in the technical sense of benefiting the group, but not an individual within a group, then $\beta_1 > 0$ and $\beta_2 < 0$.

There are several crucial implications of the expression (*). The left-hand side states that a prosocial trait z is more likely to spread when variance within groups is minimized and between-group variance is maximized. The right-hand side tells us that the strength of selection pressures is also crucial. If the trait z is extremely harmful to an individual within a group, i.e. $|\beta_2|$ is large, then it is unlikely to evolve. If it is extremely beneficial for the group, i.e. β_1 is large, then it is more likely to evolve.

This framework allows for a rigorous analysis of the sorts of design features that will enable group-beneficial adaptations, even when such adaptations run against lower-level selective pressures.

The Price equation also suggests another useful formula that can aid in understanding what conditions must be met for a group beneficial rule to proliferate.[18] As a first step, consider breaking up an individual's fitness, w_i, into two components, one determined by the rule itself and the second determined by the amount of rule adherence within the individual's group or, more precisely,*network of interaction*. Each of these components will make some separate contribution, but they are not statistically independent. Having a high level of trait z may, for example, predict that one's network of interaction is more likely to exhibit high average levels of z. This is not only because an individual will directly contribute to the average level of z within his or her group, but also because individuals with high levels of z may preferentially interact with others who exhibit a high (or low) level of z. To isolate the effects of individual z levels from those of group z levels, we must therefore write out w_i as a sum of *partial regression coefficients* (Allen, 1997). To denote the fitness effect of increasing an individual's level of rule adherence, z_i, while *holding group adherence constant*, we write the partial regression coefficient $\beta_{w_i z_i \cdot z_j}$. Similarly, to denote the fitness effect of increasing group adherence, z_j, while *holding individual adherence constant*, we write the partial regression coefficient $\beta_{w_i z_j \cdot z_i}$. Putting these together in a regression equation yields...

$$w_i = \beta_0 + \beta_{w_i z_i \cdot z_j} x_i + \beta_{w_i z_j \cdot z_i} z_j + \epsilon \qquad (5)$$

where β_0 represents base fitness and ϵ is an error term. Substituting (12) into the the Price equation produces the following equation:[19]

$$\begin{aligned} \bar{w}\Delta\bar{z} &= \beta_{w_i z_i \cdot z_j} Var(z_i) + \beta_{w_i z_j \cdot z_i} \beta_{z_j z_i} Var(z_i) \\ &= (\beta_{w_i z_i \cdot z_j} + \beta_{z_j z_i} \beta_{w_i z_j \cdot z_i}) Var(z_i) \qquad (\square) \end{aligned}$$

Since $Var(z_i) \geq 0$ as a mathematical fact, this implies that a trait will be selected for, i.e. $\bar{w}\Delta\bar{z} \geq 0$, only if we have:

$$\beta_{w_i z_i \cdot z_j} + \beta_{z_j z_i} \beta_{w_i z_j \cdot z_i} > 0 \qquad (**)$$

If we assume we are talking about an altruistic trait, then we know the following facts:

$$\beta_{w_i z_i \cdot z_j} < 0$$
$$\beta_{w_i z_j \cdot z_i} > 0$$

Given these two facts, our prediction of whether an altruistic trait will spread depends upon a crucial feature of the population-interaction structure. In order to ensure that condition (**) is satisfied, we would like $\beta_{z_j z_i}$ to be large and positive. In other words, going back to the intuitive meaning of the expression, *an altruistic trait is more likely to be selected when altruists are capable of bunching together.* This point is of fundamental importance for understanding multilevel selection and the evolution of altruism, so it bears repeating: in order for an altruistic trait to evolve (through selection), altruists must have some mechanism(s) for excluding egoists from their network of interaction or, equivalently, of converting egoists within their network into altruists.[20]

Summing up the implications of (*) and (**), we can identify four *functionlaity desiderata*:[21]

1. Prosocial traits are more likely to emerge when within-group variance $Var(z_{ij})$ is minimized, perhaps due to the ability of altruists to group together and to exclude or convert egoists.

[18] This second formula is inspired by Henrich (2004), but my formulation differs slightly. See Appendix A.4 for the derivation.

[19] See Appendix A.4 for a thorough explanation and derivation of this equation.

[20] Although, as Okasha (2006, 194) points out, this claim is true with respect to "strong altruism," but not to "weak altruism."

[21] Because these terms are important only relative to one another, these four desiderata could be reduced to two desiderata – or even to a single desideratum if we wanted to be fully parsimonious. I separate them here for analytic clarity, but we must remember that each of the desiderata must be appended with a *ceteris paribus* clause. I thank Vlad Tarko for pointing this out to me.

2. Prosocial traits are more likely to emerge when within group selective pressures against the trait ($|\beta_2|$) are small. Prosocial traits that do not require extreme sacrifice are thus more likely to proliferate, but this might also involve setting up institutional mechanisms to punish egoists and to reward altruists.
3. Prosocial traits are more likely to emerge when between group selective pressures for altruism (β_1) are large. In times of frequent and intense interaction, especially, some believe, when resources are scarce and interactions are agonistic, high levels of cooperation at the social level are imperative.
4. Finally, prosocial traits are more likely to emerge when variance between groups $Var(z_j)$ is large.

Returning to Buchanan's discussion of group-beneficial social norms, these four conditions offer a way of assuaging the fear that the "free-rider logic would seem to apply" to norms such as the work ethic (Buchanan, 1994, 81). Buchanan reasonably fears that individuals will under-invest in encouraging beneficial social norms, since these involve positive externalities.[22] However, when groups of hard-workers are able to bunch together (condition 1), forcing flower children and beach bums to also bunch together (condition 4), and when the group benefits ("positive externalities") of the work ethic are large (condition 3), while the individual cost is low (condition 2), then the Price equation implies that a strong work ethic will spread throughout the population.

In the next section, I show how these four conditions provide the key to understanding the surprising successes of polycentric governance arrangements. In short, the features of polycentric political organization help to fulfill each of the four functionality desiderata. This discussion will also address the crucial issue of monitoring and enforcement of group rules.

2.4. Group fitness and group welfare

Before applying this formal framework to the analysis of polycentric governance structures, a conceptual issue requires clarification. The Price equation tells us what will make a rule adaptive at the global level. Rules that raise the average fitness of individuals within a group will spread to individuals in other groups via imitation, immigration, conflict, or some combination of forces. The question at issue, however, is why polycentric competition often provides good governance, i.e., why it often promotes human welfare. But is there any reason to suppose that group fitness corresponds to group welfare?

Many have argued forcefully against this supposition. James Buchanan, for instance, has accused F.A. Hayek of adhering to the Panglossian fantasy that whatever evolves must be desirable. "My basic criticism of F.A. Hayek's profound interpretation of modern history and his diagnoses for improvement is directed at his apparent belief or faith that social evolution will, in fact, insure the survival of efficient institutional forms" (Buchanan, 1975, 211). Buchanan holds that Hayek's position amounts to *normative evolutionism*, that is, that we should passively accept the outcome of any evolutionary process (Buchanan, 2001, 312). In a similar vein, Dan Dennett has criticized a host of normative evolutionists for failing to explain why evolutionary outcomes should correspond to consciously chosen human values (Dennett, 1995, 468). Would survival of one's culture "justify mass murder, for instance, or betraying all your friends?" Dennett asks. Clearly not. So, why, then, should we associate evolutionary success with normative desirability?[23]

Certainly, there are ways in which norms with greater relative fitness can spread at the expense of human welfare, e.g. norms that promote violent conquest and ideological indoctrination of other groups. Such norms may be good at spreading themselves despite the fact that they are unpleasant for everyone, including members of the group in which they prevail. Just as obviously, however, a rule can attain greater relative fitness in ways that enhance group welfare, e.g. through providing cultural or economic advantages. To determine which traits will have a relative advantage requires specifying the *mechanism of selection*. If selection occurs through the ability to excel at violence, or to impose other costs on competing groups, then it is quite likely that the fitness of a rule will not correspond to its ability to enhance welfare.[24] If, on the other hand, selection occurs through the ability to induce others to adopt one's own rules due to their intrinsic appeal or their apparent ability to promote the well-being of the group that adheres to them, then it is quite likely that the most fit rules will also be those that promote human welfare.

In the context of cultural evolution within a polycentric framework, there are at least three reasons to think that group fitness does, in fact, correspond to group welfare. These reasons allow us to discount the worries put forth by Buchanan, Dennett, and others.

First, although these worries are quite troubling for the case of genetic evolution, a major criterion of selection in *cultural* evolution is the appeal of certain rules or norms. In the case of genetic evolution, new adaptations are introduced randomly, and their ability to proliferate is determined purely in terms of their ability to promote relatively more offspring than alternative genetic sequences. By contrast, as explained above (sec. 2.2), *cultural* fitness is determined, at least partly, by the

[22] Buchanan (1994, 70) uses the phrase "paying the preacher" to refer to any investment in spreading and encouraging work ethic.

[23] One response, put forth by Hayek on various occasions, is that our deepest values are themselves products of cultural and biological evolution. For instance, Hayek writes that "value ... can only be understood as the determinant of what people must do to maintain the overall structure" (Hayek, 1983, 36). This response raises a host of difficulties and complexities, which, if left unaddressed, render it unconvincing. Although I believe there is some merit to this response, laying it out in sufficient detail would lead us far afield.

[24] Bowles and Gintis (2011, ch.8, 197) emphasize the effect of violent conflict on cultural selection, while Mesoudi (2011) emphasizes the ability of rules to benefit their adherents as a source of fitness.

appeal of a particular rule of conduct.[25] As Ostrom (1999, 57-8) points out, institutional evolution is not entirely blind, but is partly directed by the goal of improving human welfare. Institutional changes often involve copying rules that individuals know will improve their well-being. Alternatively, in the absence of such detailed knowledge, individuals may simply copy the rules of societies that appear to be highly successful, even if they do not entirely understand why (Henrich, 2015).[26]

A second, related reason to think that group fitness will correspond to group welfare in the case of cultural evolution concerns the variety of positive sum ways in a rule can exhibit a high degree of cultural fitness. In the case of both cultural and genetic evolution, one way of increasing the relative fitness of a unit of selection is by reducing the number of copies produced by other units of selection. In the case of genetic selection, this generally involves imposing a welfare cost on other organisms, since shortening their lives or decreasing their resources is the simplest way to limit their reproductive success. In the case of cultural selection, by contrast, decreasing the fitness of another norm need not involve harming the welfare of the organisms adhering to it. One can make a norm less appealing simply by providing better alternatives, or perhaps by "reframing" the norm is a way that supports a more negative attitude towards that norm (Bicchieri, 2016, 121-2, 126, 139).

Although it is less likely in the case of cultural evolution than in the case of genetic evolution, sometimes the best way to make a norm unappealing *is* by imposing costs on those individuals or groups who hold that norm. Here the specific case under consideration, viz. cultural evolution *within a polycentric framework*, offers a third reason for connecting group fitness with group welfare. As elaborated below, polycentric orders are governed by an overarching set of rules (Tarko, 2021).[27] Such rules generally aim to reduce non-productive, zero-sum competition between groups, while, at the same time, promoting rivalry that leads to useful institutional experimentation. To be "fit" within such a framework, a rule must offer apparent benefits, since the overarching set of rules limits the margins along which it can actively reduce the welfare of other individuals within the system. In this way, the overarching set of rules offers a framework for cultural evolution that supports a correlation between evolutionary fitness and human welfare. In some cases, the overarching set of rules will be dysfunctional. Due to this possibility, polycentricity alone does not suffice to ensure beneficial outcomes. Nevertheless, as argued in the next section, it does exhibit several desirable properties, which greatly increase the likelihood of beneficial evolutionary outcomes.

3. Polycentricity and multilevel selection

3.1. Defining polycentricity

Vincent Ostrom offers a concise and now classic definition of polycentricity:

> ...a polycentric political system [is] composed of: (1) many autonomous units formally independent of one another, (2) choosing to act in ways that take account of others, (3) through processes of cooperation, competition, conflict, and conflict resolution. (V. Ostrom 1991, 225)

Other scholars have built upon this definition to provide greater precision. Especially notable are two teams of scholars. First, Aligica and Tarko (2012, 257) present a "concept design" that involves three key features of polycentricity, as well as a host of empirical indicators for each of these features. The three features are:

1. A multiplicity of decision centers
2. An overarching system of rules
3. A process of evolutionary competition between the decision centers.

A second team of scholars—Stephan, Marshall, and McGinnis (2019)—provide a list of eight features of polycentric systems. However, they consider four of these features to be of special importance:

1. Multiple decision centers
2. Autonomous decision-making authority for each decision center
3. Overlapping jurisdictions of authority between the decision centers
4. Various processes of mutual adjustment among decision centers (41).

While several others have also provided definitions of polycentricity, they all more or less resemble the three definitions covered here.[28]

Pulling together the various features of these three different definitions, we can identify a basic schema for the organization of polycentric governance:

[25] 'Rule of conduct' is used loosely here. The same reasoning applies to any other sort of cultural replicator.

[26] Ostrom and Henrich also point out that cultural evolution takes place on much faster time scale than does biological evolution, largely due to the fact that it incorporates deliberate choice and intentional modifications, rather than random, blind variation. I thank an anonymous referee for suggesting the importance of this point.

[27] Absent an overarching set of rules, the system is not technically polycentric, but fragmented and anarchic (Tarko, 2016, 43).

[28] For alternative, though similar, definitions, see Ostrom et al. (1961), Toonen (1983), Folke et al. (2005). Aligica and Boettke (2009), Garmestani and Benson (2013).

Polycentric Political Structure: A *polycentric political structure* consists of rule-governed collectives with well-defined, and often overlapping jurisdictions that interact in a rule-governed, competitive manner resulting in the relative expansion or contraction of their jurisdictions.

To reduce this definition to an orderly list, we might identify three features:

1. Multiple decision-making units
2. Each decision-making unit has authority over a specified, but evolving jurisdiction
3. These decision-making units compete and cooperate in various ways, according to a set of overarching rules.

This precise understanding of polycentricity enables an assessment of polycentric functionality in terms of the conditions for group functionality laid out in Section 2.3, an assessment to which we now turn.

3.2. Facilitating multilevel selection

The key to understanding how polycentric political organization enables beneficial outcomes is to appreciate how the features of polycentricity, as defined above, coincide with the conditions for group-beneficial outcomes derived in Section 2.3. To begin, recall the first two functionality desiderata, both of which concern the *within-group* components of evolutionary pressures:

1. Prosocial traits are more likely to emerge when within-group variance $Var(z_{ij})$ is minimized.
2. Prosocial traits are more likely to emerge when within group selective pressures against the trait ($|\beta_2|$) are small.

Polycentric organization facilitates the satisfaction of (1) by allowing like-minded individuals to coalesce into groups centered around shared concerns. The importance of self-sorting for the maintenance of cooperation has long been recognized by social scientists. As Axelrod (1986, 1105) writes in his classic article on the evolution of norms, an important "mechanism for the support of norms is voluntary membership in a group working together for a common end." This is also a common point raised in favor of federalist political constitutions, a form of polycentric political organization.[29] In a federalist structure, "[m]obile individuals can join that city-state having their most preferred set of rights and responsibilities" (Inman and Rubinfeld, 1996). More recently, the fact that polycentric orders provide a framework for voluntary coalitions has been emphasized by Aligica (2018), who argues that polycentric political structures enable the spontaneous formation of groups in which individuals coalesce around a defined "problem solving" context Aligica (2018, 104). Having agreed on the existence and nature of a pressing problem, citizens are more likely to coordinate on shared rules that seek to address it. These features of a problem solving context provide reason to believe that individuals within the group are more likely to adhere to the prevailing social rules. Within-group variance, in other words, is minimized by the spontaneous formation of like-minded groups. And a polycentric political framework facilitates this formation.

Axelrod and Aligica also believe that voluntary group formation supports effective *monitoring and punishment*, the presence of which directly addresses functionality desideratum (2).

> The power of membership works in three ways. First, it directly affects the individual's utility function, making a defection less attractive because to defect against a voluntarily accepted commitment would tend to lower one's self-esteem. Second, group membership allows like-minded people to interact with each other, and this self-selection tends to make it much easier for the members to enforce the norm implicit in the agreement to form or join a group. Finally, the very agreement to form a group helps define what is expected of the participants, thereby clarifying when a defection occurs and when a punishment is called for (Axelrod, 1986, 1105-6).

Having formed a like-minded group, with general agreement on priorities, citizens are more likely to accept compromises, do their fair share, and censure violators Aligica (2018, 104). Importantly, violators themselves are more likely to respond positively to censure, since they accept the basis of social rules and recognize their importance for solving a relevant problem.[30]

There is general agreement among evolutionary theorists that effective monitoring and punishment can stabilize cooperative behavior. It does so by removing the advantages of rule-breaking, thus minimizing $|\beta_2|$. However, there is also widespread agreement that punishment itself is costly:

> It might be argued that individuals cooperate in order to avoid punishment by other members of their own group. This notion seems plausible based on common experience. However, it does not solve the theoretical problem; it only raises the new problem of why individuals should cooperate to punish other individuals. Punishment itself is an investment in the production of some other public good, for example, civil order. Each potential punisher can have

[29] Jan Vogler has convinced me that not *all* federalist systems are truly polycentric. For instance, there may not be cooperation and competition between political units on the same level. That said, most federalist systems do fall into the category of polycentric organizations.

[30] Whether or not the explanation offered by Axelrod and Aligica is found to be convincing, there is ample empirical evidence that cooperation increases when individuals are allowed to enter or exit groups facing social dilemmas. See, for example, Orbell and Dawes (1993), Orbell et al. (1984), Schuessler (1989), and Yamagishi and Hayashi (1996). Whatever the explanation, self-sorting appears to be a powerful mechanism for stabilizing cooperation.

> only a small incremental effect on the level of civil order, and again, the cost to the individual participating in the punishment of others could be substantial. The rational selfish individual would let the other person do the punishing (Boyd and Richerson, 1982, 238)[31]

While the first-order problem of cooperation may be resolved by punishment norms, we still face a "second-order social dilemma (of equal or greater difficulty)" (Ostrom, 1998, 7). This is no trivial matter, since it concerns the evolutionary stability of group-beneficial rules (Dawkins, 2016).[32] To complete the account of group functionality, then, we must consider whether polycentric governance structures have the capacity to support effective monitoring and punishment.

A wide range of recent work on the evolution of punishment supports the idea that punishment norms, despite their altruistic character, can become widespread and stable within properly structured populations. For our purposes, there are two key ideas in this literature. The first is that punishment exhibits decreasing costs as it becomes more widespread. As Bowles and Gintis (2011, 149) put it, "punishment is characterized by increasing returns to scale, so the total cost of punishing a particular target declines as the number of punishers increases." In fact, to the extent that punishment is effective at suppressing norm violations, "... the payoff disadvantage of punishers relative to contributors approaches zero as defectors become rare because there is no need for punishment" (Boyd et al., 2003, 3531). In other words, once punishment is widespread and effective within a group, there is little to no fitness differential between punishers and non-punishers. So, if the punishment norm is the trait in question, then $|\beta_2|$ will be extremely low, permitting group selection to overpower individual-level selection in the spread of punishment norms (Boyd et al., 2003, 3534).

The idea of decreasing punishment costs does not, however, explain how punishment might arise in the first place. To explain the emergence, rather than the mere stability, of punishment norms requires the second key idea: self-sorting. In general, cooperative behaviors are far more stable when cooperators are able to identify themselves and group together (Eshel and Cavalli-Sforza, 1982). In the case of punishment, the increasing returns to scale noted by Bowles and Gintis (2011, 149), make self-sorting all the more effective.[33] Boyd et al. (2010) develop a model in which individuals can signal to one another that they are committed to the same norms and are willing to enforce them by punishing defectors. This ability to form coalitions with like-minded individuals reduces the risk that one will suffer major harms by attempting to unilaterally punish defectors. Self-sorting thus supports the spread of punishment norms by allowing for coordinated punishment efforts. In this scenario, therefore, punishment norms can proliferate even when they start off at extremely low levels. Moreover, the logic of this model provides theoretical intuition for the empirical finding that individuals are more willing to engage in costly punishment when they are able to communicate and thereby coordinate their punishing behavior, a phenomenon observed in the laboratory (Ostrom et al., 1992, 405), as well as in the field (Ostrom, 2014, 244).

The many articles cited above suggest that self-sorting will support effective punishment and thereby reduce $|\beta_2|$. The key lesson, for our purposes, is that punishment norms can proliferate and stabilize cooperation when willing punishers are able to sort themselves into groups. When the network of interaction supports non-random encounters, so that cooperators and willing punishers are more likely to interact, then punishment can become common within the group and functionality desideratum (2) will be satisfied. As argued above, *the capacity to self-sort into a community with shared priorities or interests is a key design feature of polycentric organizations*. For this reason, polycentric governance supports effective punishment norms and hence the satisfaction of functionality desideratum (2).

The second two functionality desiderata concern the group-level forces:

3. Prosocial rules are more likely to emerge when between group selective pressures (β_1) are large.
4. Prosocial rules are more likely to emerge when variance between groups $Var(z_j)$ is large.

The third feature in our definition of a polycentric political structure – that decision-making units interact and compete – favors the increase of β_1. When jurisdictions with diverse rule sets are in a state of constant interaction, members of other jurisdictions become familiar with the alternative rules and with their effects with respect to well-being. This, in turn, heightens inter-unit competition. If the unit's jurisdiction is geographical, they are more likely to engage in *Tiebout competition* by enticing individuals to "vote with their feet" (Tiebout, 1956). If the jurisdiction is non-geographical, individuals may simply switch to the governance provider who yields better results at lower costs. In this way, interaction between jurisdictions intensifies group-level selective pressures by increasing competition in the standard way familiar to all economists. Rules that fail to entice adherents thus face more rapid decline than they would under autarkic conditions with lower levels of competition.

The first and second features in our definition of a polycentric political structure – multiple decision-making units with authority over a specified jurisdiction – favor prosociality by increasing $Var(z_j)$. Boundaries between groups are crucial for developing distinctive sets of rules; institutional diversity presupposes distinct jurisdictions.[34] Without well-defined groups, in which individuals share the same rules, the variance between collectives will either be less pronounced or undefined,

[31] See also Ostrom (1998, 7).

[32] For the original arguments, see Williams (1966).

[33] The increasing returns dynamic is also discussed by Boyd et al. (2003, 3531), Henrich (2004, 26-7), Henrich and Boyd (2001, 81), and Boyd et al. (2010, 617-8).

[34] It's important to bear in mind that jurisdictions may or may not be territorial. Religion, for example, has been analyzed as a polycentric order and its jurisdictions are defined by its members, not by a geographical area (Gill, 2020).

since the collectives themselves will be undefined. The capacity for polycentric systems to support institutional diversity has been elaborated by Aligica (2014): "a polycentric system is the embodiment of institutional pluralism....The pluralism of institutional forms ensures a variety of responses to a variety of circumstances....[O]ne can hardly think of a better arena for experimentation than polycentricity" (66). As Aligica points out, polycentric organization offers an institutional framework in which groups or coalitions can experiment with a wide array of different rules to respond to a wide variety of different concerns. The result is that different groups within society will form around different sets of rules, thus increasing the diversity between groups and satisfying functionality desideratum (4).

In sum, polycentric governance structures directly address all four functionality desiderata. The aim has not been to show that polycentricty will guarantee that any of these desiderata will be satisfied. Instead, the aim has been to demonstrate that polycentric structures support these desiderata, thus increasing the *likelihood* that the social system will satisfy them. When these desiderata are all satisfied, group-level selection will overpower individual-level selection. In such a case, the outcome will be higher levels of social functionality. The Price equation framework thus offers a theoretical tool for understanding how polycentric governance structures can be effective even though they lack a centrally-orchestrated plan for achieving beneficial outcomes.

3.3. Polycentricity in action

The discussion thus far has been highly abstract. To clarify the theory expounded here, it is worth considering a concrete example of polycentric governance in action. A classic example of polycentric governance is the set of institutions, comprising formal and informal rules, that regulates the scientific community.[35] Polanyi (1951), who originally articulated the concept of polycentricity, presents science, as practiced in countries like the United Kingdom, as the paradigmatic example of polycentric organization. More recent work has confirmed and deepened Polanyi's analysis (Tarko, 2015). This subsection will show how science fits the definition of polycentricity and demonstrate that this allows it to fulfill functionality desiderata (1)–(4).

To understand the polycentric organization of science, it helps to contrast it with an example of non-polycentric science. At the time when Polanyi was writing on polycentricity, the Soviet Union was engaged in a large-scale planning experiment. Their monocentric governance approach extended to science, as well, where the Soviet government laid down methods, rules, and doctrines for scientific researchers. This was especially true in the area of research on genetics, where Trofim Denisovich Lysenko, director of the USSR's Institute of Genetics, persecuted scientists who pursued research in areas that he deemed to be pseudo-scientific or corrupted by bourgeois ideas. Most notably, Lysenko rejected Mendelian genetics, and, consequently, outlawed research in this area. Polanyi (1951, 107) cites Lysenko's ordered execution of Nikolai Vavilov, but there are several other notable examples, such as the geneticist Nikolai Balyaev.[36] In short, scientists who did not adhere to the official state views on science were punished severely by officials of the centralized state.

This monocentric organizational scheme stands in stark contrast to the polycentric scheme that prevails in free nations around the world. Scientific progress occurs as parallel research teams – often employing different methods and accepting different theoretical premises – compete for publications, citations, and awards. Though diverse in many ways, these different teams generally accept a thin set of professional norms, including a commitment to seeking truth and to publicizing their findings and evidence (Smolin, 2006, 301). As Tarko (2015) explains, these features of the scientific community suffice to make it a form of polycentric organization. The decentralized teams constitute a set of decision-making units, satisfying the first feature of polycentricity as laid out in Section 3.1. In addition, decision-making units have broad autonomy in determining the kind of research they will carry out, along with the kind of research they will valorize by citing or building upon. This realm of authority, sometimes physically housed in formal institutions, such as universities or government-funded research centers, constitutes a specified jurisdiction, thus satisfying the second feature of polycentric organization. Finally, these research teams compete in various ways: to attract funding, to publish papers, to garner citations, and to win awards. They also cooperate by sharing data, further developing one anothers' theories, or offering constructive criticism at professional conferences. These various ways of interacting conform to a set of overarching rules, a general "shared ethic" (Smolin, 2006, 301), that allows them to improve the state of science through their interactions. The organization of the scientific community thus satisfies the third feature of polycentric organization, as well.

The theory developed here concerns the ability of a polycentric organizations to achieve successful outcomes, and in the case of science, this success is spectacular:

> The scientific community is arguably one of the most successful human organizations ever created, both with respect to its declared main purpose (truth-seeking) and with respect to secondary goals such as obtaining large government subsidies (while maintaining independence and freedom from interference) and obtaining preferential treatment in public schools or in courts of law (despite often being highly disruptive to common belief systems) (Tarko, 2015, 64)

[35] Other real-world examples include certain metropolitan governance systems (Ostrom et al., 1961), political federations, such as the European Union (Vogler, 2020), and decentralized, competitive resource management schemes, such as the CAMPFIRE project in Zimbabwe (Schmidtz, 1997).

[36] Dmitri Balyaev, the brother of Nikolai, went on to discover several genetic principles behind domestication, but was forced to do so in secret, due to his fear of upsetting party officials (Wrangham, 2019, 67ff.).

What I have called "the puzzle of polycentric functionality" emerges from an appreciation of what can go wrong when individuals engage in unplanned, decentralized activities. As Polanyi (1951, 108) puts it, "Suppose we started building a house without any plans, each workman adding his part according to his own ideas, using whatever materials he preferred, putting in bricks or timber, lead pipes or floorboards as he thought fit. Surely the result would be a hopeless confusion." In response to this puzzle, Polanyi (1951, 109) asserts that the "nature of scientific systems is more akin to the ordered arrangement of living cells which constitute a polycellular organism" than it is to the construction of a home. In his own intuitive way, Polanyi has therefore suggested that the success of science can be understood in terms of multilevel selection. The transition from unicellular to multicellular organisms is accounted for by increasing pressures at the cellular-group level and decreasing pressures at the individual cellular level (Smith and Szathmary, 1997). Polanyi's analogy therefore evokes exactly the type of explanation provided here, although his statement anticipates by twenty years the formal analysis of multilevel selection in the Price equation framework.

Following the lead of Polanyi, we can use the framework developed above to explain how the polycentric organization of science allows researchers to operate like cells in a body. The polycentric nature of science means that there is no monopoly provider of research; researchers can self-sort into teams of like-minded researchers, intent on employing similar methods to pursue knowledge of similar topics. Desideratum (1), which requires reducing within-group variance, is naturally satisfied in this way.

Satisfying desideratum (2), which requires the minimization of within-group selective pressures, is not terribly difficult in a context where the team members share a common fate. Individuals who refuse to accept the norms (i.e. the premises and methods) of their research team will likely disrupt the process of research and publication, hurting themselves in the process. This is, therefore, an example of "inclusive fitness," which theorists have shown to be conducive to promoting altruistic traits (Hamilton, 1964).[37] There will likely be cases where the joint nature of success is still insufficient to suppress rogue scientists who threaten to hurt the team's progress or lazy scientists who free ride on the efforts of others. Although far less severe than Soviet-style punishment, research teams possess their own battery of punitive measures for researchers who reject the prevalent norms of their institution. These may include refusing to include a scientist on publications and, in the academic context, refusing to grant tenure. In addition, such scientists may simply be ignored: "[t]he ultimate fate of the entrant who disagrees with the orthodoxy but cannot persuade the community to accept his point of view is, quite simply, isolation within or banishment from the community" (Kendall, 1960, 979). Even highly prestigious individuals who refuse to accept the norms of their research team or their broader community wind up languishing in isolation. As Tarko (2015, 69) explains, this was the fate of Albert Einstein when he rejected the prevailing interpretation of quantum mechanics.

Desideratum (3) concerns the strength of between-group selective pressures. In the scientific community, these pressures are exerted in a decentralized fashion. They include social prestige, funding, academic titles, and awards. They are powerful motivators, since they determine the career success and reputation of those who devote their lives to research. Important, in this respect, is that research teams typically sink or swim together, and teams regularly view themselves as competing with other teams. One of the most well-known examples of a highly competitive interaction between research teams is the race which unfolded throughout the 1950s to discover how the various parts of the DNA molecule fit together. Two leading teams emerged: one team, led by Watson and Crick, were in a conscious effort to outpace that led by Wilkins and Franklin. The result was a major scientific breakthrough. Successful teams, as the DNA example underlines, become highly influential in the field of science, and their "norms" – that is, their basic premises and methods – are copied by other research teams hoping to make their own contributions and acquire their own prestige.

Finally, desideratum (4), which requires diversity between groups, may be the most prominent benefit of decentralized scientific organization. Although there are areas of relatively settled science, in which an overwhelming consensus prevails, on many issues disagreement is quite common. In such areas, pluralism is crucial, as it allows "as many trails as possible [to] be covered" (Polanyi, 1951, 110). As Tarko (2015, 71) explains,

> As long as there are grounds for reasonable people to disagree, the polycentric nature of the scientific community is crucial for its success because it is this polycentric organization that secures the diversity of opinions. It is not enough to rely on individual scientists being creative and able to "think outside of the box." It is essential for them to have institutional environments where they can pursue their viewpoints.

The polycentric organization of science provides such institutional environments, and in this way supports a greater diversity between research units than a monocentric system like that implemented by the Soviet Union.

In sum, the polycentric organization of scientific inquiry supports the conditions required for promoting cooperative norms and suppressing opportunistic behavior. In this regard, it greatly outperforms more monocentric alternatives. The main thesis of this paper is that the success of polycentric organizations can be explained by their capacity to promote higher-level selection pressures and suppress lower-level ones. The search for truth undertaken by the highly polycentric scientific community demonstrates how this capacity operates in real world institutional settings.

[37] More precisely, from the replicator's perspective, these traits cease to be altruistic, since harming others entails harming oneself. Norms that permit this will not proliferate.

4. Conclusion

Drawing on the theory of multilevel selection, this paper has addressed the question of why polycentric political structures yield successful outcomes. Polycentricity allows a system to satisfy four functionality desiderata derived from the multilevel Price equation. These desiderata represent the conditions under which group benefits exert more force relative to individual benefits. Polycentric organizations structure human interacitons so as to increase the likelihood of satisfying these desiderata. Through the prism of polycentricity, individual behaviors are synthesized into group-level adaptations.

It is important to emphasize, however, that polycentric structures do not ensure group-level success. In particular, mechanisms to ensure effective monitoring and enforcement may not develop, even though polycentric structures increase the likelihood of such mechanisms. From a theoretical perspective, the Price equation suggests that polycentric governance requires some way of reducing the within-group benefit of rule-violating behavior. Empirical case studies confirm this theoretical insight: several of Elinor Ostrom's design principles can be understood precisely in terms of the need to reduce the benefits of rule violation, and her case studies demonstrate how failure ensues when a self-governing community fails to establish effective monitoring and punishment (Ostrom, 2016, Ch. 5). Group functionality can hardly emerge amidst ubiquitous defection and free-riding. As argued in Section 3.2, however, work on the evolution of punishment suggests that polycentric structures support, even if they don't guarantee, the development of effective punishment norms.

This paper has sought to create a conceptual bridge between Ostromian political economy and multilevel selection theory. Certain features of polycentric institutional structures – i.e. well-defined, competing jurisdictions with rule-making autonomy – provide the requisite conditions for group-level adaptation to take place. In a polycentric structure, group-beneficial norms, including norms of monitoring and enforcement, are likely to spread.

A crucial question, however, remains unanswered: how do we explain the emergence of polycentric governance structures themselves? Addressing this question presents an opportunity to deepen our understanding of polycentricity along both historical and theoretical lines. Must polycentric institutions be designed and imposed or can they themselves evolve spontaneously? If so, under what conditions should we expect such institutions to evolve? Because polycentric institutions are themselves frameworks for group-beneficial adaptations, these questions are related to exciting work on the evolution of evolvability (Wagner and Altenberg, 1996).

Declaration of Competing Interest

None.

Data availability

No data was used for the research described in the article.

Appendix A

George Price wanted a simple, but extremely general equation to describe the change in the average level of some character trait between "generations." The word "generations" is in quotations, since Price's equation describes even processes that take place without any genetic relatedness, indeed, without any genes whatsoever. The Price equation thus describes technological or cultural evolution just as well as genetic evolution.

A1. Deriving price's equation

A1.1. Set-up

- There is a population P consisting of n entities.
  - 'P' stands for 'Parent'.
- z: A (phenotypic) trait of some kind, which can be measured (discretely or continuously) with real numbers.
- z_i: The level of z exhibited by the entity $i \in P = \{1, \ldots, n\}$.
- $\bar{z}$: The average level of z in the P-population.
  - Mathematically, $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$
- z_i': The amount of trait z transmitted by i to any of its "offspring."
- Δz_i: The change in the level of z from one entity to its offspring.
  - $\Delta z_i = z_i' - z_i$
  - z_i' can be thought of as the *copying fidelity* or *transmission bias* of trait z for entity i.
  - Taking an average across all entities in P, the expectation of Δz_i is $\mathbf{E}[\Delta z] = \frac{1}{n}\sum_{i=1}^{n} \Delta z_i$
- w_i: The average number of offspring produced by entity i.
  - w_i can be thought of as the "fitness" of entity i.
  - Similarly to $\bar{z}$, we define the average fitness as $\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$
  - Often we will be interested, not in the absolute fitness of an entity, but in its fitness *relative* to the other entities in the group. For this comparative purpose, we define *relative fitness* of entity i as $\gamma_i = \frac{w_i}{\bar{w}}$.

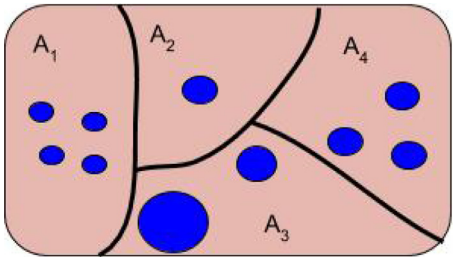

Fig. A.1. The Partition Theorem Illustrated by a Truly Great Artist.

- There is another population of interest, that comprised of the offspring of all entities in *P*. Call this population *O*.
  - '*O*' stands for 'Offspring'.
- $\bar{z}_o$: The average level of the z trait in population *O*.
  - Mathematically,

$$\bar{z}_o = \frac{1}{n}\sum_{i=1}^{n}\gamma_i z_i' = \frac{1}{n}\sum_{i=1}^{n}\frac{w_i}{\bar{w}}z_i' \tag{6}$$

  - Importantly, for the derivation, $\mathbf{E}[\gamma] = \frac{1}{n}\sum_{i=1}^{n}\gamma_i = 1$, which makes sense given that γ_i is a proportion of the total *O*-population.

The Partition Theorem Equation (6), while it describes the *O*-population, is couched entirely in terms of *P*-population traits. This may seem somewhat mysterious, but it is the most crucial equation to understand for the derivation that follows. This equation is easy to grasp once one understands the so-called *partition theorem.*[38] The idea is actually quite intuitive. Suppose we divide the total population into a set of jointly exhaustive and mutually exclusive subpopulations. Then the average level of a trait, e.g. z, within the population as a whole will simply be the (weighted) sum of the averages of each subpopulation, where each one is weighted by its relative size. This idea can be illustrated with a simple diagram.

In Fig. A.1, we have divided a population comprised of dots into four subpopulations, $A_1 - A_4$. We might think of each of these subpopulations as the offspring of a single entity in the *P*-population, so that A_1 consists entirely of entity 1's offspring, A_2 of entity 2's, and so on. Suppose the size of each dot represents the magnitude of our quantity of interest (here, the amount of trait z each entity in the *O*-population has). The partition theorem states that we should add up the average z level in each of the subpopulations $A_1 - A_4$, weighting each of these by the proportion of the total entities that they contain. Here, A_1 has many entities, $4/10 = 40\%$ of them, but these entities don't possess very much of the z-trait. So, they will drag the average down much more than, say, A_2, which has a modest amount of z-trait and only comprises $1/10 = 10\%$ of the total population. A_3 will surely raise the average, but not by too much, since it only contains 2 entities, 20% of the population. A_4, on the other hand, contains a sizeable 30% of the total population, but the average amount of z-trait in A_4 seems neither large nor small. Hence, adding A_4 to the calculation is unlikely to significantly raise or lower the running average. Letting $\bar{A}_j$ denote the average level of z-trait in subpopulation $j \in \{1, 2, 3, 4\}$, the partition theorem tells us that the average (i.e. the expected value) of the whole population will be:

$$\frac{4}{10}\bar{A}_1 + \frac{1}{10}\bar{A}_2 + \frac{2}{10}\bar{A}_3 + \frac{3}{10}\bar{A}_4.$$

The reasoning applied here is exactly the reasoning that underlies the partition theorem, and if you understood this reasoning, then you are (at least) very close to understanding why $\bar{z}_o = \frac{1}{n}\sum_{i=1}^{n}\gamma_i z_i'$.

A1.2. Deriving $\Delta\bar{z}$

Again, what we're after is a simple expression that describes the change in the average level of the z-trait. We will denote this quantity $\Delta\bar{z}$. Now, to begin this derivation, we simply note the fact that $\Delta\bar{z}$ must be equal to the difference between the average in the new *O*-population, and the old *P*-population.

$$\begin{aligned}\Delta\bar{z} &= \bar{z}_o - \bar{z} \\ &= \frac{1}{n}\sum_{i=1}^{n}\gamma_i z_i' - \frac{1}{n}\sum_{i=1}^{n} z_i\end{aligned}$$

[38] See any probability textbook for an explanation. A particularly nice statement, couched in terms of expected values, rather than probabilities, can be found in Grimmett and Welsh (2014, 34).

$$\text{Recall,} \qquad z_i' = \Delta z_i + z_i$$

$$\begin{aligned}\text{Hence,} \qquad \Delta\bar{z} &= \frac{1}{n}\sum_{i=1}^{n}\gamma_i(\Delta z_i + z_i) - \frac{1}{n}\sum_{i=1}^{n} z_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\gamma_i z_i - \frac{1}{n}\sum_{i=1}^{n} z_i + \frac{1}{n}\sum_{i=1}^{n}\gamma_i\Delta z_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\gamma_i z_i - \mathbf{E}[\gamma]\frac{1}{n}\sum_{i=1}^{n} z_i + \frac{1}{n}\sum_{i=1}^{n}\gamma_i\Delta z_i \\ &= \mathbf{E}[\gamma z] - \mathbf{E}[\gamma]\mathbf{E}[z] + \mathbf{E}[\gamma\,\Delta z] \qquad (*)\end{aligned}$$

Covariance The covariance between two random variables, X and Y with mean values μ_X and μ_Y, respectively, is defined

$$Cov(X,Y) = \mathbf{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

Therefore,

$$\Delta\bar{z} = Cov(\gamma, z) + \mathbf{E}[\gamma\,\Delta z]. \qquad \text{(Price Eq.)}$$

For some reason, presumably mathematical tractability, biologists typically prefer to multiply (Price Eq.) through by $\bar{w}$, yielding an equivalent, but more familiar, statement of the Price equation:

$$\bar{w}\Delta\bar{z} = Cov(w, z) + \mathbf{E}[w\Delta z] \qquad (2)$$

Notice that the only real change is that γ has been replaced with w. This is just because $\gamma_i = w_i/\bar{w}$, and we multiplied through by $\bar{w}$. The mathematical details here are both simple and unenlightening, so I will spare the reader. To obtain 2, simply multiply equation $(*)$ by $\bar{w}$ and simplify.[39]

A2. Deriving the multilevel price equation

The key idea of the multilevel version of the Price equation is to partition selective pressures into two categories: one between "collectives" and one between "particles," i.e. the elements which make up the collectives. Building from Eq. (2), there are two ways of arriving at the multilevel from of the Price equation. The top-down approach relies on the neat recursive trick of inserting the Price equation into itself, but indexing the inserted version to a lower level of selection. The bottom-up approach also relies on a mathematical trick, that of decomposing covariance into a within-partition term and a between-partition term.

A2.1. Top-Down

Recall Eq. (2):

$$\bar{w}\Delta\bar{z} = Cov(w, z) + \mathbf{E}[w\Delta z].$$

The covariance and expectation terms of this equation are taken over entities $i \in P$. In this top-down version of the derivation, we will assume that each entity i is itself a collective. That is, each $i \in P$ is itself made up of adaptive particles. We might think of each i as a group made up of individual organisms, or as an organism made up of genes which are sometimes capable of manipulating the meiotic process so as to increase their own spread, often at the expense of the progeny-organisms (meiotic drive).

To formalize this new situation, we will need some additional notation. Let $j \in i$ be particles in group i. For ease of mathematics, and without loss of generality, we assume that all groups i are of the same size. In this context, we will slightly abuse notation to draw an important distinction: $Cov_i(w, z)$ will represent the covariance between group-fitness (i.e. the average number of offspring produced by the particles in i) and group-trait level (i.e. the average level of z possessed by the particles in i). $Cov_j(w, z)$, on the other hand, will represent the covariance between particle-fitness (i.e. the number of offspring produced by each $j \in i$) and group-trait level (i.e. the level of z possessed by the $j \in i$) *within some group i*. Similarly, $\mathbf{E}_i[w\Delta z]$ will represent the expectation of the product of group-fitness (i.e. the average number of offspring produced by the particles in i multiplied by the change in group-trait level (i.e. the change in average level of z possessed by the particles in i between generations). $\mathbf{E}_j[w\Delta z]$, on the other hand, will represent the expectation of particle-fitness (i.e. the number of offspring produced by each $j \in i$) multiplied by the change in the individuals' levels of z-trait (i.e. the level of z possessed by the $j \in i$). More briefly, when the expectation is indexed by i, the expectation sums across the groups i that partition P. When the expectation is indexed by j, the expectation sums across particles j that comprise some specified group i. And similarly for covariance.

[39] To see a derivation of (2) instead of (Price Eq.), see Okasha (2006, ch.1).

We will derive the following:

Lemma. In the multilevel context, Eq. (2) is equivalent to the following:

$$\bar{w}\Delta\bar{z} = Cov_i(w,z) + \mathbf{E}_i\left[Cov_j(w,z) + \mathbf{E}_j[w\Delta z]\right] \tag{7}$$

Proof. Starting with Eq. (2) and inserting our new notation...

$$\begin{aligned}\bar{w}\Delta\bar{z} &= Cov(w,z) + \mathbf{E}[w\Delta z] \\ &= Cov_i(w,z) + \mathbf{E}_i[w\Delta z] \qquad (\mathrm{i}-\mathrm{level})\end{aligned}$$

That is, we apply the Price equation at the level of collectives, made up of lower level particles that are not yet represented in the equation. We are moving in a "top-down" direction. Let m be the number of particles in each group and n be the number of groups. To represent the lower-level particles explicitly, we note that...

$$\begin{aligned}\mathbf{E}_i[w\Delta z] &= \mathbf{E}\left[\frac{1}{m}\sum_{j=1}^{m} w_j \frac{1}{m}\sum_{j=1}^{m}\Delta z_j\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} w_j \frac{1}{m}\sum_{j=1}^{m}\Delta z_j\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\bar{w}_j\Delta\bar{z}_j) \\ &= \mathbf{E}_i\left[\bar{w}_j\Delta\bar{z}_j\right]\end{aligned}$$

Here, we note that the expression within the expectation identical to the left-hand side of Eq. (2), except that it is indexed to the lower level of particles j, rather than the level of collectives i. Accordingly, we can pull the clever move I alluded to above, of recursively inserting the Price equation into itself:

$$\begin{aligned}\mathbf{E}_i[w\Delta z] &= \mathbf{E}_i\left[\bar{w}_j\Delta\bar{z}_j\right] \\ &= \mathbf{E}_i\left[Cov_j(w,z) + \mathbf{E}_j[w\Delta z]\right] \qquad (\mathrm{j}-\mathrm{level})\end{aligned}$$

We now simply insert (j-level) into the expression (i-level) to yield our result:

$$(7)\ \bar{w}\Delta\bar{z} = Cov_i(w,z) + \mathbf{E}_i\left[Cov_j(w,z) + \mathbf{E}_j[w\Delta z]\right]$$

□

Before re-deriving (7) from the bottom up, consider its interpretation. This equation partitions the selective forces into two levels: the i-level of collectives, represented by $Cov_i(w,z)$ and the j-level of particles, represented by $\mathbf{E}_i\left[Cov_j(w,z) + \mathbf{E}_j[w\Delta z]\right]$. What would happen if we eliminated evolutionary pressures at the particle level by assuming that all particles "breed true" ($z' = z$ and hence $\mathbf{E}_j[w\Delta z] = 0$) and that they all have the same fitness (so that z_j is not correlated with w_j)? We are left only with $Cov_i(w,z)$, which measures how collective-level fitness (the average fitness of a collective's particles) corresponds to collective-level z-trait (the average z-trait level of a collective's particles). In other words, we are left with selection at the level of collectives. Alternatively, suppose that there is no collective-level selection. That is, a higher average level of the z-trait within a group does not correspond to greater proliferation of its members relative to other groups. Perhaps all collectives have the same fitness, or, for whatever other reason, w_i is not correlated with z_i. Then, of course, $Cov(w_i, z_i) = 0$ and we are left only with the term $\mathbf{E}_i\left[Cov_j(w,z) + \mathbf{E}_j[w\Delta z]\right]$, which takes the average across groups of the evolutionary outcomes that occur within groups at the particle level. In other words, the evolutionary process is entirely determined by selection and transmission bias at the particle-level.

One more point, which will serve as a transition to bottom-up thinking, bears mentioning. The term $\mathbf{E}_i\left[Cov_j(w,z) + \mathbf{E}_j[w\Delta z]\right]$ has been described in two distinct ways: (1) as selection and transmission bias when seen from the particle-level, and (2) as the transmission bias when seen from the collective level. The Price equation thus reveals an interesting fact about multilevel selection: transmission bias at level i is an evolutionary process unto itself at level $(i-1)$. To take a concrete example, if we observe the preferential transmission of a particular allele in a human population, this can be understood either as a sort of transmission bias at the phenotypic level ($z_i' > z_i$ and so $\mathbf{E}[w_i\Delta z_i] > 0$), or we can consider it to be a process of meiotic drive, an evolutionary process unto itself, taking place at the level of competing genes. As we will see shortly, starting at either of these levels, we can build upwards to consider how lower-level evolutionary processes determine evolution at higher levels. Evolving systems, at least in the Price framework, can be seen as a vast series of nested and inter-determining levels of selection.

A2.2. Bottom-Up

Just as the top-down derivation employed a clever recursive trick, so this bottom-up derivation utilizes another tool in the arsenal of mathematical analysis: decomposing covariance into a within- and a between-group component. To demonstrate this decomposition I rely on graphics and intuition. A fully rigorous mathematical approach can be found in Wade (1985, 62-3).

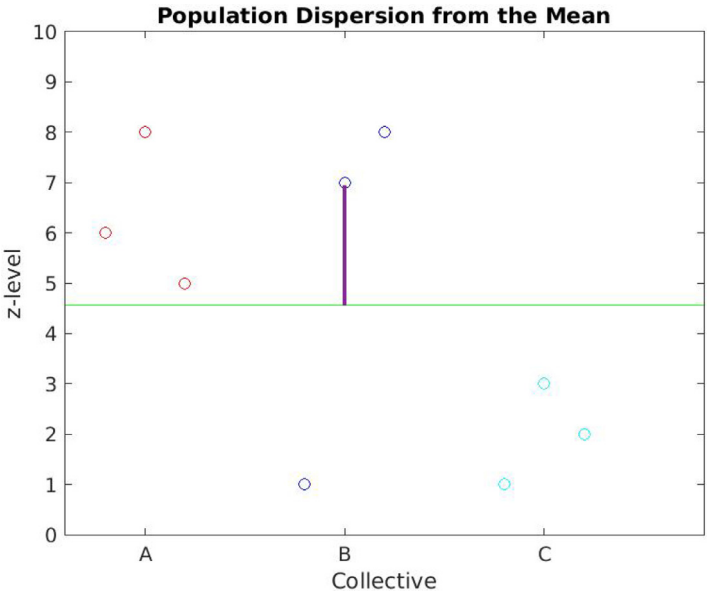

Fig. A2. Population Dispersion of z-trait.

Consider the covariance term we wish to decompose: $Cov(w, z)$. Our decomposition involves separating an observed dispersion into two components: one that is "explainable" by the group to which the particles belong and another explainable by the dispersion within the group. Suppose we have a total population P consisting of 9 particles, all of whom have varying levels of z-trait. There are 9 total particles, $j \in \{1 \ldots, 9\}$, but considered as members of a group, each particle will be indexed with an $i \in \{1, 2, 3\}$. We thus have three groups, $k \in \{A, B, C\}$ (Fig. A.2).

Each circle marker in this graph represents a particle j in the population P. In this figure, the fact that they are grouped into three distinct collectives is irrelevant. We are considering the total dispersion of the particles, measured by the sum of their distance from the population mean, represented by the horizontal green line. The center particle, directly above 'B', shows how we measure this quantity: for each particle, we draw a line like the one connecting the center particle to the mean. Then we sum up this different across all particles. Now, this would be a poor measure of dispersion, since extremely above-average and extremely below average particles would cancel out, making the dispersion seem small when it is really large. For that reason, these difference are typically squared to yield the formula for mean-squared distance:

$$MSD = \sum_{j=1}^{9} (z_j - \bar{z})^2,$$

where z_j is as above, $\bar{z}$ is the population average z-level (here, 4.66).[40]

Our formula will be rather different, however, since we are not interested in variance, but in *covariance*. So, we must construct a second graph, similar to the first, except on the y-axis we have fitness w, instead of z. Then, instead of MSD, we will calculate:

$$Cov(w, z) = \frac{1}{9} \sum_{j=1}^{9} (w_j - \bar{w})(z_j - \bar{z}).$$

The second graph corresponds to the first term in the covariance expression, $(w_j - \bar{w})$, which measures our fitness dispersion (Fig. A.3):

Again, we calculate the dispersion by subtracting each j's value from the population mean. Our expression for covariance now conveys useful information: if w_j tends to move in the same direction as z_j, then $Cov(w, z)$ will be fairly large. If, on the other hand, they tend to move in opposite directions, then $Cov(w, z)$ will be negative. If they exhibit little to no correspondence, then $Cov(w, z)$ will be close to 0.

Notice, further, that we have represented z as a fairly altruistic trait: within groups, those with higher z-level have fitness that is lower than the *group* average. Consider for example, $j = 2$ with a high $z_2 = 8$, but a fitness level w_2 that is lower than

[40] The sum is often multiplied by $\frac{1}{n}$ to give us an average, denoted S^2, rather than a total that strictly increases as we add more particles. When used for estimation purposes, as in many statistics textbooks, the sum is divided by $n - 1$ rather than n for somewhat esoteric reasons involving the desire for an unbiased estimator.

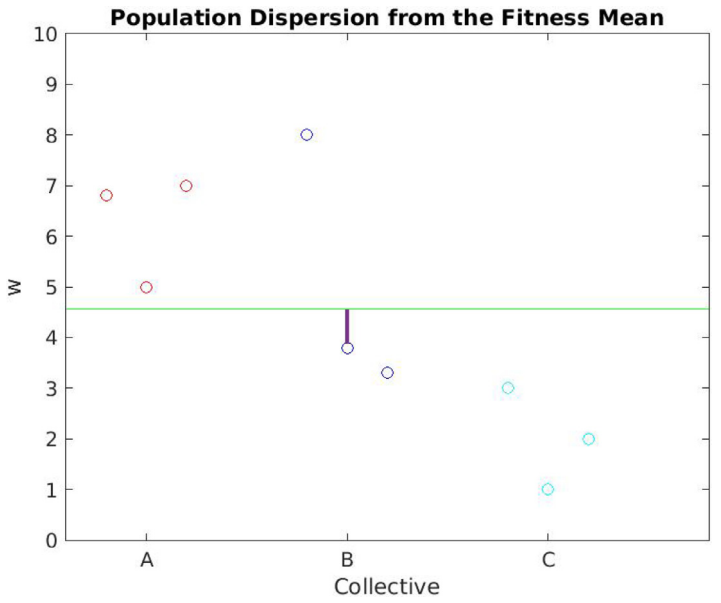

Fig. A3. Population Dispersion of Fitness.

either of its two collective members in A. Nevertheless, group A as a whole does quite well. Even the highly self-sacrificial particle $j = 2$ outperforms all members of the selfish group C.

Now, what the decomposition technique shows is that we can break up these dispersion-measures into within- and between-group components. This is easily visualized. First consider between-collective dispersion (Fig. A.4).

Now the centered vertical line measures the distance between the B-collective mean and the population average, representing the center particle indirectly, only through its influence on the collective mean (the thick blue bar). In practice, we draw a similar graph for fitness w, calculating the distance between each collective's mean and the population mean. Then we calculate the between-group covariance of z-level and w, denoting this value $Cov_k(w, z)$. We also run through a similar process to calculate *within*-group covariance between z-level and w.

In Fig. A.5, we add up the differences between particles' z-levels and the *collective* average, rather than the population average. Hence, we draw a line from the center particle to 5.33, its within-collective average, rather than to 4.66, the total population average. Again, we do something similar to find the dispersion of within-group fitness w. Now, clearly, differences within the group account for much of the total population dispersion. But, just as clearly, they do not account for all of it, because the within-collective averages are (almost by definition) closer to their within-group particles than they are to the whole population-wide gamut of particles. The rest of the population dispersion is captured by the dispersion of collective means from the population mean, visualized in Fig. A.4. The claim of the decomposition technique is that the total population dispersion can be captured by summing the *average* within-collective covariance and the between-collective covariance.

With this intuition, let's briefly formalize the decomposition claim. If we let i index particles within collectives k, then w_{ik} is the ith particle in the kth group, with $i \in \{1, 2, 3\}$ and $k \in \{A, B, C\}$. For example, w_{2B} denotes the fitness of the center particle, i.e. the second particle in the B group. Similarly, z_{ik} is the z-level of the ith particle in the kth group. If we write simply w_k, then we are denoting the within-collective mean of k. As above, $Cov_k(w, z)$ denotes the covariance between the within-collective average fitness and the within collective average z-level, expanded this is: $Cov_k(w, z) = \frac{1}{3}\sum_{k=A}^{C}(w_k - \bar{w})(z_k - \bar{z})$. On the other hand, we will use Cov_i to denote a covariance within a group k, so that $Cov_i(w, z) = \frac{1}{3}\sum_{i=1}^{3}(w_{ik} - \bar{w}_k)(z_{ik} - \bar{z}_k)$. With this notation, we can formally spell out the intuitive claim that total population dispersion is decomposable into between- and within-group components (Fig. A.5):

$$Cov(w, z) = Cov_k(w, z) + \mathbf{E}_k[Cov_i(w, z)] \tag{8}$$

The population-covariance between fitness w and z-trait is equal to the sum of (i) the covariance between collective fitness w_k and collective z-trait (i.e. the averages within the collective) and (ii) the mean of the within-collective covariances between fitness w and z-trait.

Armed with Eq. (8), the bottom-up derivation of Eq. (7) is quite simple.

Lemma. In the multilevel context, Eq. (2) is equivalent to the following:

$$\bar{w}\Delta\bar{z} = Cov_k(w, z) + \mathbf{E}_k[Cov_i(w, z) + \mathbf{E}_i[w\Delta z]]$$

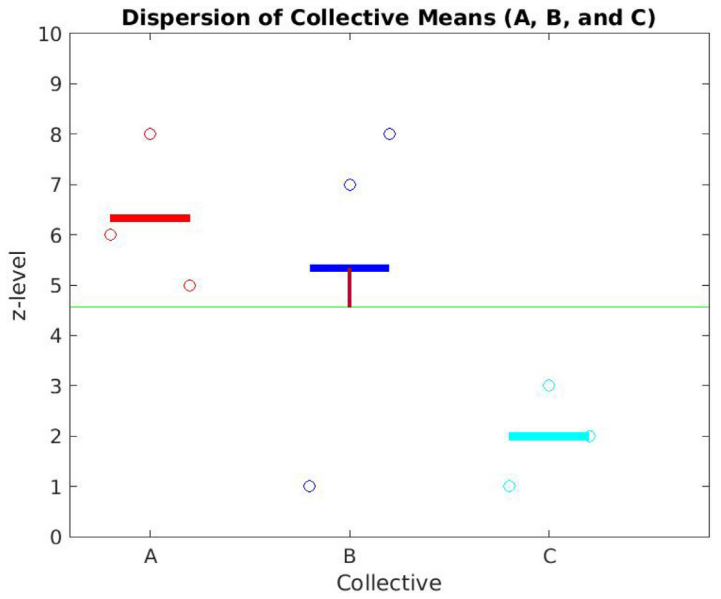

Fig. A4. z-Dispersion Between Groups.

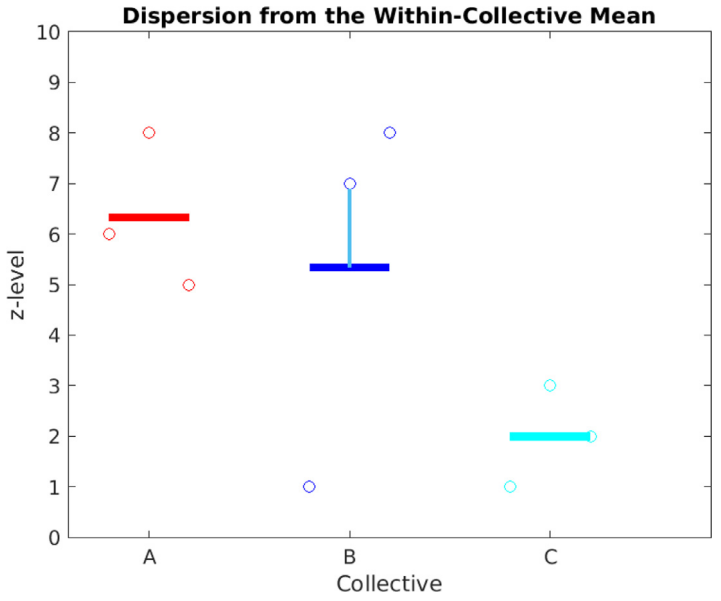

Fig. A5. z-Dispersion Within Groups.

Proof. Recall Eq. (2):

$$\begin{aligned}\bar{w}\Delta\bar{z} &= Cov(w,z) + \mathbf{E}[w\Delta z] \\ &= Cov_k(w,z) + \mathbf{E}_k[Cov_i(w,z)] + \mathbf{E}[w\Delta z]\end{aligned}$$

An easily derivable fact about expected value is that, for any partition of the population P, the expectation of the total population is equal to the expected value of the sum of the expected values of the partitions. Hence,

$$\mathbf{E}[w\Delta z] = \mathbf{E}_k[\mathbf{E}_i[w\Delta z]] = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_K}\sum_{i=1}^{n_K}(w_{ik}\Delta z_{ik}),$$

where K is the total number of collectives (elements in the partition), n_K is the total number of particles in collective K, w_{ik} is the fitness level of the ith particle in the kth collective, and z_{ik} is the z-level of the ith particle in the kth collective.

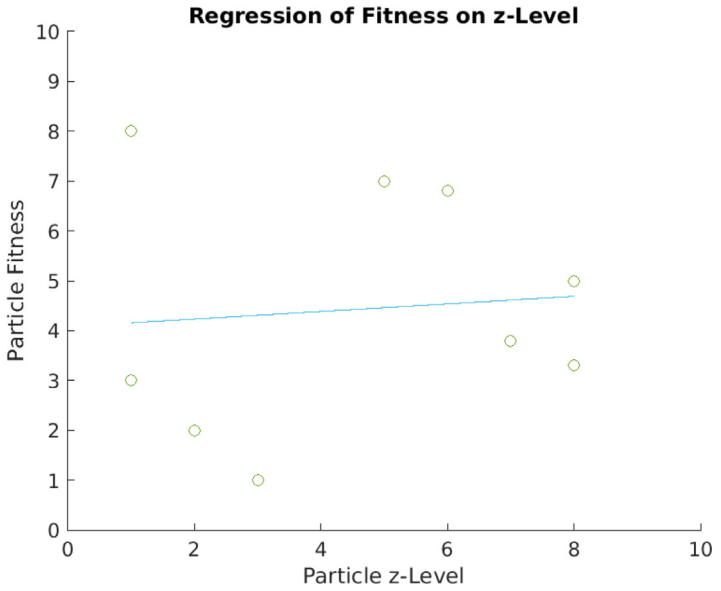

Fig. A6. Regression of w on z.

Inserting this equation into the above expression...

$$\begin{aligned}\bar{w}\Delta\bar{z} &= Cov_k(w,z) + \mathbf{E}_k[Cov_i(w,z)] + \mathbf{E}_k[\mathbf{E}_i[w\Delta z]] \\ &= Cov_k(w,z) + \mathbf{E}_k[Cov_i(w,z) + \mathbf{E}_i[w\Delta z]]\end{aligned}$$

□

In contrast to the top-down derivation, this one began with particles, not considering whether these particles are themselves collectives comprised of lower level particles. They may be. We could then apply the same top-down derivation to these particles, identifying them as intermediary collectives, made up of particles but also constituting particles for higher-level collectives. Evolutionary systems are thus seen to be nested processes, teeming with activity at each level, an activity that often expresses itself through emergent coherence and unity at higher levels.

A3. Linear regression

Linear Regression Given a sample of data points, a linear regression is a line of best fit, one which minimizes the summed distance of points from the line.[41] The equation for this line of best fit is represented as...

$$w = \beta_0 + \beta_1 z + \epsilon, \tag{9}$$

where w is the variable we are running the regression *of* and z is the variable this regression is *on*. β_0 and β_1 are the constants chosen to minimize the summed distance of data points from the line. Since the model will not be perfect, ϵ is an error term, which we will assume exhibits no systematic bias (i.e. the mean $\mathbf{E}[\epsilon] = 0$).

We can visualize this line with the data we posited above when analyzing the within- and between-collective components of covariance (Fig. A.6):

This figure plots the various particles from the population discussed above and fits a line that minimizes the summed distance between the particles and this line. In this case, we get $\beta_0 = 4.081$, $\beta_1 = .076$, and hence $w = .076z + 4.081$. This tells us that there is a slight positive correlation between fitness and the z trait, despite that fact that this trait requires self-sacrifice on the part of the particles that possesses it. In other words, selection at the collective-level is slightly overpowering selection at the particle-level in this population.

But how did we calculate β_0 and β_1? There are several methods (and computer programs), but the most instructive is to consider the expected values of our variables and to solve for β_0 and β_1.

$$\begin{aligned}w &= \beta_0 + \beta_1 z + \epsilon \\ \mathbf{E}[w] &= \mathbf{E}[\beta_0 + \beta_1 z + \epsilon] \\ &= \mathbf{E}[\beta_0] + \mathbf{E}[\beta_1 z] + \mathbf{E}[\epsilon]\end{aligned}$$

[41] Where distance is measured by the square of the difference.

$$= \beta_0 + \beta_1 \mathbf{E}[z] + 0$$

Therefore, $\beta_0 = \mathbf{E}[w] - \beta_1 \mathbf{E}[z]$

Next, consider the covariance between w and z:

$$\begin{aligned} Cov(z, w) &= Cov(z, \beta_0 + \beta_1 z + \epsilon) \\ &= Cov(z, \beta_0) + Cov(z, \beta_1 z) + Cov(z, \epsilon) \qquad \text{(From properties of coviariance)} \\ &= \beta_0 Cov(z, 1) + \beta_1 Cov(z, z) + Cov(z, \epsilon) \\ &= 0 + \beta_1 Var(z) + 0 \end{aligned}$$

This last step follows from the independence of z, ϵ, and the fact that $Cov(X, X) = Var(X)$.

We therefore conclude:

$$\beta_0 = \mathbf{E}[w] - \beta_1 \mathbf{E}[z] \qquad \beta_1 = \frac{Cov(z, w)}{Var(z)} \tag{10}$$

The second equation in (10) shows how we derive the equations that begin Section 2.3.

A4. The price equation and partial regression coefficients

Recall the basic form of the Price Equation:

$$\bar{w}\Delta\bar{z} = Cov(w_i, z_i) + \mathbf{E}[w_i \Delta z] \tag{11}$$

Now, let us consider fitness of individual i, w_i, as depending partly on the trait z_i and partly on the average level of the altruistic train within the group, denoted z_j. Given that it is an altruistic trait, fitness will be undermined by the possession of this trait, but enhanced by others possessing it. And since i's possession of the trait will be correlated with the group's possession of this trait, we need a way to separate out these two effects. A simple regression equation can do this for us. Let $\beta_{w_i z_i \cdot z_j}$ be the partial regression coefficient for w_i on z_i. That is, $\beta_{w_i z_i \cdot z_j}$ represents the strength of the effect that changing z_i has on w_i *holding* z_j *constant*. Similarly, $\beta_{w_i z_j \cdot z_j}$ will represent the partial regression coefficient for w_i on z_j. That is, $\beta_{w_i z_j \cdot z_i}$ represents the strength of the effect that changing z_j has on w_i *holding* z_i *constant*. As is standard, we will also include the y-intercept term β_0 and an error term ϵ:

$$w_i = \beta_0 + \beta_{w_i z_i \cdot z_j} z_i + \beta_{w_i z_j \cdot z_i} z_j + \epsilon \tag{12}$$

Now, take the expression in 12 and substitute it into expression 11. Let $\Delta z = 0$, since we are not interested in drift, but selection. So, setting $\mathbf{E}[w_i \Delta z] = 0$ and substituting...

$$\bar{w}\Delta\bar{z} = Cov(\beta_0 + \beta_{w_i z_i \cdot z_j} z_i + \beta_{w_i z_j \cdot z_i} z_j + \epsilon, z_i) \tag{13}$$

Relevant Covariance Rules

$$\begin{aligned} Cov(X + Y, Z) &= Cov(X, Z) + Cov(Y, Z) \\ Cov(aX, Z) &= aCov(X, Z) \\ Cov(X, X) &= Var(X) \\ Cov(X, Y) &= \beta_{YX} Var(X) \end{aligned}$$

Applying covariance rules to Eq. (13), we get:

$$\bar{w}\Delta\bar{z} = \beta_{w_i z_i \cdot z_j} Var(z_i) + \beta_{w_i z_j \cdot z_i} Cov(z_j, z_i)$$

Because $Cov(z_j, z_i) = Cov(z_i, z_j) = \beta_{z_j z_i} Var(z_i)$...

$$\begin{aligned} \bar{w}\Delta\bar{z} &= \beta_{w_i z_i \cdot z_j} Var(z_i) + \beta_{w_i z_j \cdot z_i} \beta_{z_j z_i} Var(z_i) \\ &= (\beta_{w_i z_i \cdot z_j} + \beta_{z_j z_i} \beta_{w_i z_j \cdot z_i}) Var(z_i) \qquad (\square) \end{aligned}$$

To understand the meaning of this last expression, we can borrow from the analysis of Frank (1998) to reconstruct the complicated coeffiecient term in ($\square$). Let the regression coefficient $\beta_{w_i z_i}$ denote exactly what we want to know: the total, non-decomposed, statistical (i.e. not necessarily causal) effect of a trait z_i on the organism's fitness, w_i. Frank points out that this regression coefficient contains two separable components. One is the direct, causal effect of z_i on w_i, ignoring any indirect effects that result from the group. We have been representing this component with the familiar partial regression coefficient $\beta_{w_i z_i \cdot z_j}$. The second component consists of (i) the statistical effect of trait z_i on the group-level of trait z, denoted z_j, multiplied by (ii) the causal effect of the group level z_j on an organism's fitness w_i. We denote (i) with the regression coefficient $\beta_{z_j z_i}$ and (ii) with the partial regression coefficient $\beta_{w_i z_j \cdot z_i}$. This can be visualized by the following diagram, adapted from Frank (1998, 52) (Fig. A.7):

As this diagram indicates...

$$(\boxtimes) \beta_{w_i z_i} = \beta_{w_i z_i \cdot z_j} + \beta_{z_j z_i} \beta_{w_i z_j \cdot z_i}$$

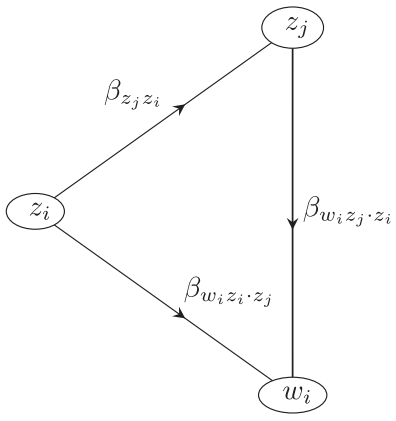

Fig. A7. Relating partial regression coefficients and fitness.

Comparing (□) and (⊠), we see that the right hand side of (⊠) is equal to the coefficient in front of $Var(z_i)$ in (□). This is fantastic, because we now have some intuitive sense of what the the expression (□) actually signifies. The adaptiveness, or tendency to increase, of a trait z is regulated partly by the effect it has on the organism's fitness, but also partly by the product of (i) how likely an organism with that trait is to be surrounded by other organisms with that trait and (ii) the impact on the organism of being in a group with the trait z.[42] It's nice to understand what the mathematics means, but we should also ask what we can do with it.

Henrich (2004) is able to get some serious mileage out of the expression (□). Since $Var(z_i) \geq 0$ as a mathematical fact, this implies that a trait will be selected for, i.e. $\bar{w}\Delta\bar{z} \geq 0$, only if we have:

$$\beta_{w_i z_i \cdot z_j} + \beta_{z_j z_i}\beta_{w_i z_j \cdot z_i} > 0 \qquad (**)$$

If we assume we are talking about an altruistic trait, then we know the following facts:

$$\beta_{w_i z_i \cdot z_j} < 0$$
$$\beta_{w_i z_j \cdot z_i} > 0$$

Given these two facts, our prediction of whether an altruistic trait will spread depends upon a crucial structural feature of the population. In order to ensure that condition (∗∗) is satisfied, we would like $\beta_{z_j z_i}$ to be large and positive. In other words, going back to the intuitive meaning of the expression, *an altruistic trait is more likely to be selected when altruists are capable of bunching together.* This point is of fundamental importance for understanding multilevel selection and the evolution of altruism, so it bears repeating: in order for an altruistic trait to evolve (through selection), altruists must have some mechanism(s) for excluding egoists from their network of interaction or, equivalently, of converting egoists within their network into altruists.

References

Aligica, P.D., 2014. Institutional Diversity and Political Economy: The Ostroms and Beyond. Oxford University Press.
Aligica, P.D., 2018. Public Entrepreneurship, Citizenship, and Self-Governance. Cambridge University Press.
Aligica, P.D., Boettke, P.J., 2009. Challenging Institutional Analysis and Development: The Bloomington school. Routledge.
Aligica, P.D., Tarko, V., 2012. Polycentricity: from Polanyi to Ostrom, and Beyond. Governance 25 (2), 237–262.
Allen, M.P., 1997. Partial regression and residualized variables. In: Understanding Regression Analysis, pp. 86–90.
Andersson, K.P., Ostrom, E., 2008. Analyzing decentralized resource regimes from a polycentric perspective. Policy Sci. 41 (1), 71–93.
Anomaly, J., Brennan, G., 2014. Social norms, the invisible hand, and the law. U. Queensland LJ 33, 263.
Axelrod, R., 1986. An evolutionary approach to norms. Am. Polit. Sci. Rev. 80 (4), 1095–1111.
Bicchieri, C., 2016. Norms in the Wild: How to Diagnose, Measure, and Change Social Norms. Oxford University Press.
Bowles, S., Gintis, H., 2011. A Cooperative Species. Princeton University Press.
Boyd, R., Gintis, H., Bowles, S., 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science 328 (5978), 617–620.
Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proc. Natl. Acad. Sci. 100 (6), 3531–3535.
Boyd, R., Richerson, P.J., 1982. Cultural transmission and the evolution of cooperative behavior. Hum. Ecol. 325–351.
Buchanan, J., 1994. Ethics and Economic Progress. University of Oklahoma Press.
Buchanan, J.M., 1975. The Limits of Liberty: Between Anarchy and Leviathan. Liberty Fund.
Buchanan, J.M., 2001. Federalism, Liberty, and the Law. Liberty Fund.
Carlisle, K., Gruby, R.L., 2019. Polycentric systems of governance: a theoretical model for the commons. Policy Stud. J. 47 (4), 927–952.
Dawkins, R., 2016. The Selfish Gene: 40th Anniversary Edition. Oxford University Press, New York, NY.
Dennett, D.C., 1995. Darwin's Dangerous Idea: Evolution and the Meanings of Life. Simon and Schuster.
Eshel, I., Cavalli-Sforza, L.L., 1982. Assortment of encounters and evolution of cooperativeness. Proc. Natl. Acad. Sci. 79 (4), 1331–1335.
Folke, C., Hahn, T., Olsson, P., Norberg, J., 2005. Adaptive governance of social-ecological systems. Annu. Rev. Environ. Resour. 30, 441–473.

[42] Again, remember that, although we are treating z as a binary trait for ease of expression, the same reasoning applies to a continuous trait.

Frank, S.A., 1998. Foundations of Social Evolution. Princeton University Press.
Gardner, A., 2008. The price equation. Curr. Biol. 18 (5), R198–R202.
Garmestani, A.S., Benson, M.H., 2013. A framework for resilience-based governance of social-ecological systems. Ecol. Soc. 18 (1).
Gill, A., 2020. The comparative endurance and efficiency of religion: a public choice perspective. Public Choice 1–22.
Gintis, H., 2009. Game Theory Evolving. Princeton University Press.
Grimmett, G., Welsh, D.J.A., 2014. Probability: An Introduction, 2nd ed. Oxford University, Oxford.
Hamilton, W.D., 1964. The genetical theory of social behaviour. I, II. J. Thour. Biol. 7, 1–52.
Hayek, F.A., 1983. Knowledge, Evolution, and Society. Adam Smith Institute.
Henrich, J., 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. J. Econ. Behav. Organ. 53 (1), 3–35.
Henrich, J., 2015. The secret of our success. The Secret of Our Success. Princeton University Press.
Henrich, J., Boyd, R., 2001. Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. J. Theor. Biol. 208 (1), 79–89.
Inman, R.P., Rubinfeld, D.L., 1996. The political economy of federalism. In: Mueller, D. (Ed.), Perspectives on Public Choice. Cambridge University Press.
Kendall, W., 1960. The æopen societyg and its fallacies. Am. Polit. Sci. Rev. 54 (4), 972–979.
Martin, N.P., Storr, V.H., 2008. On perverse emergent orders. Stud. Emergent Order 1 (1), 73–91.
Mesoudi, A., 2011. Cultural evolution. Cultural evolution. University of Chicago Press.
Okasha, S., 2006. Evolution and the levels of selection. Clarendon Press; Oxford University Press, Oxford: Oxford; New York. OCLC: ocm70985413
Orbell, J.M., Dawes, R.M., 1993. Social welfare, cooperators' advantage, and the option of not playing the game. Am. Sociol. Rev. 787–800.
Orbell, J.M., Schwartz-Shea, P., Simmons, R.T., 1984. Do cooperators exit more readily than defectors? Am. Polit. Sci. Rev. 78 (1), 147–162.
Ostrom, E., 2016. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press.
Ostrom, E., 1998. A behavioral approach to the rational choice theory of collective action. Am. Polit. Sci. Rev. 92 (1), 1–22.
Ostrom, E., 1999. Coping with tragedies of the commons. Annu. Rev. Polit. Sci. 2 (1), 493–535.
Ostrom, E., 2009. Understanding Institutional Diversity. Princeton University Press.
Ostrom, E., 2014. Collective action and the evolution of social norms. J. Nat. Resour. Policy Res. 6 (4), 235–252.
Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: self-governance is possible. Am. Polit. Sci. Rev. 86 (2), 404–417.
Ostrom, V., Tiebout, C.M., Warren, R., 1961. The organization of government in metropolitan areas: a theoretical inquiry. Am. Polit. Sci. Rev. 55 (4), 831–842.
Polanyi, M., 1951. The Logic of Liberty: Reflections and Rejoinders. Liberty Fund.
Samuelson, P.A., 1954. The pure theory of public expenditure. Rev. Econ. Stat. 36 (4), 387–389.
Schmidtz, D., 1997. When preservationism doesn't preserve. Environ. Values 6 (3), 327–339.
Schuessler, R., 1989. Exit threats and cooperation under anonymity. J. Conflict Resolut. 33 (4), 728–749.
Smith, J.M., Szathmary, E., 1997. The Major Transitions in Evolution. OUP Oxford.
Smolin, L., 2006. The Trouble with Physics: The Rise of String Theory, the Fall of a Science, and What Comes Next. Mariner Books.
Stephan, M., Marshall, G., McGinnis, M., 2019. An introduction to polycentricity and governance. In: Governing complexity: analysing and applying polycentricity, pp. 21–44.
Tarko, V., 2015. Polycentric structure and informal norms: competition and coordination within the scientific community. Innov. Eur. J. Social Sci.Res. 28 (1), 63–80.
Tarko, V., 2016. Elinor Ostrom: An Intellectual Biography. Rowman & Littlefield.
Tarko, V., 2021. Polycentricity. Routledge Handbook of PPE. Routledge.
Tarko, V., 2022. Polycentricity. In: Melenovsky, C.M. (Ed.), The Routledge Handbook of Philosophy, Politics, and Economics. Routledge.
Thiel, A., Garrick, D.E., Blomquist, W.A., 2019. Governing Complexity: Analyzing and Applying Polycentricity. Cambridge University Press.
Tiebout, C.M., 1956. A pure theory of local expenditures. J. Polit. Economy 64 (5), 416–424.
Toonen, T.A.J., 1983. Administrative plurality in a unitary state: the analysis of public organisational pluralism. Policy & Politics 11 (3), 247–271.
Turchin, P., 2011. Warfare and the evolution of social complexity: a multilevel-selection approach. Struct. Dyn. 4 (3).
Vogler, J.P., 2020. The political economy of the European union: an exploration of EU institutions and governance from the perspective of polycentrism. In: Exploring the Political Economy and Social Philosophy of Vincent and Elinor Ostrom, edited by Peter J. Boettke, Bobbi Herzberg, and Brian Kogelmann, pp. 145–181.
Wade, M.J., 1985. Soft selection, hard selection, kin selection, and group selection. Am. Nat. 125 (1), 61–73. Publisher: [University of Chicago Press, American Society of Naturalists]
Wagner, G.P., Altenberg, L., 1996. Perspective: complex adaptations and the evolution of evolvability. Evolution 50 (3), 967–976.
Williams, G.C., 1966. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton University Press.
Wilson, D.S., 2016. Two meanings of complex adaptive systems. In: Complexity and Evolution: Toward a New Synthesis for Economics, pp. 31–46.
Wilson, D.S., Ostrom, E., Cox, M.E., 2013. Generalizing the core design principles for the efficacy of groups. J. Econ. Behav. Organ. 90, S21–S32.
Wrangham, R., 2019. The Goodness Paradox: The Strange Relationship Between Virtue and Violence in Human Evolution. Vintage.
Yamagishi, T., Hayashi, N., 1996. Selective play: social embeddedness of social dilemmas. In: Frontiers in Social Dilemmas Research. Springer, pp. 363–384.